|  |  |  |  |
| --- | --- | --- | --- |
| **Name (Last, First)** | **StudentID** | **URL for Word/PDF file** | **For Marking**  **(Yes/No)** |
| **Hirenkumar Patel** | **9029986** | [**Assignment 1.docx**](https://github.com/hpatel9986/Web_Analytics/tree/main/Assignment_1) | **No** |
| **Sreejith Nair** | **9045575** |  | **Yes** |

**Task-1:  
1. Create new table with the specific column name as per given task.**

Removed column from the table with the blank value and irrelevant

# Required fields

required\_fields = [

    'Object ID', 'Department', 'Object Name',

    'Title', 'Culture', 'Artist Nationality',

    'Object Begin Date', 'Object End Date',

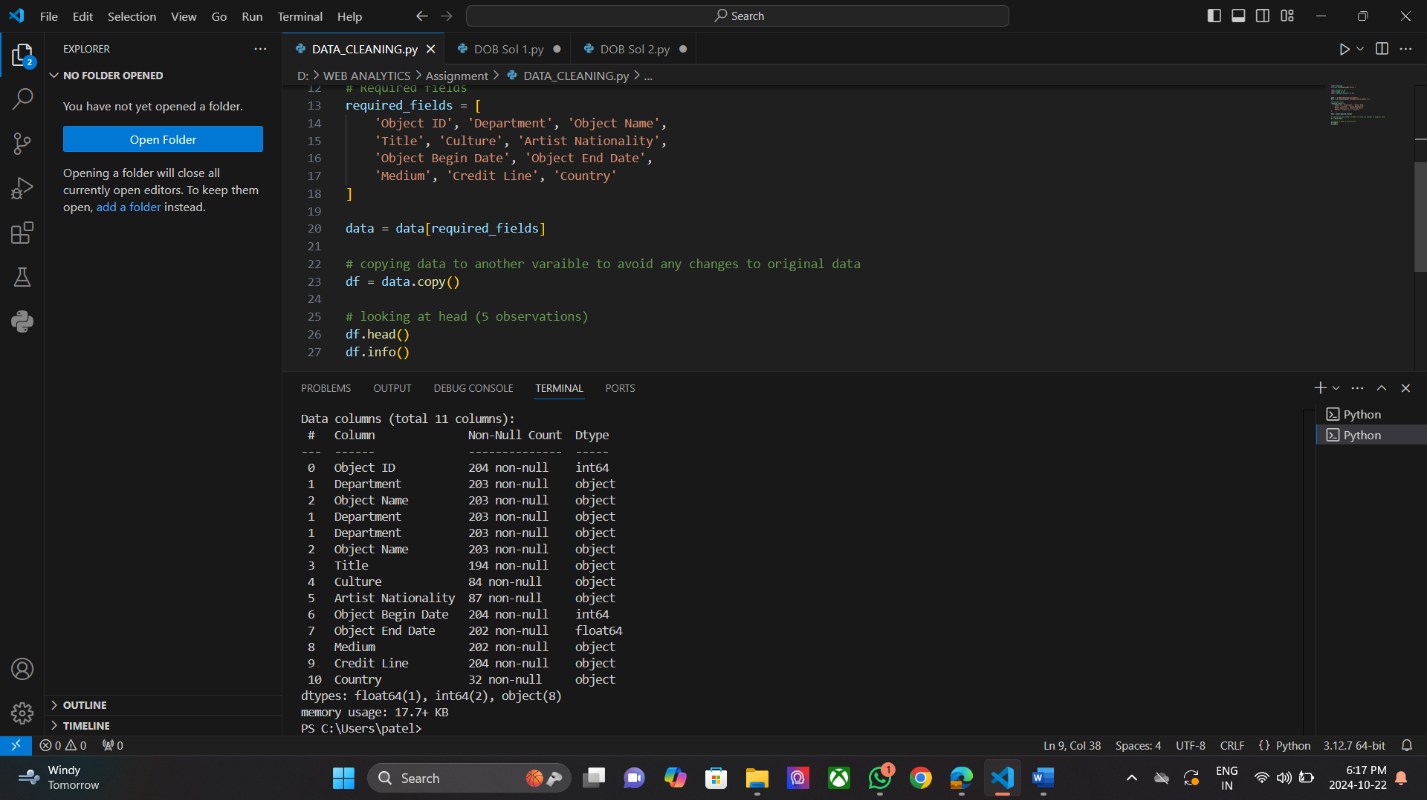
    'Medium', 'Credit Line', 'Country'

]

data = data[required\_fields]

df = data.copy()

df.info()

****

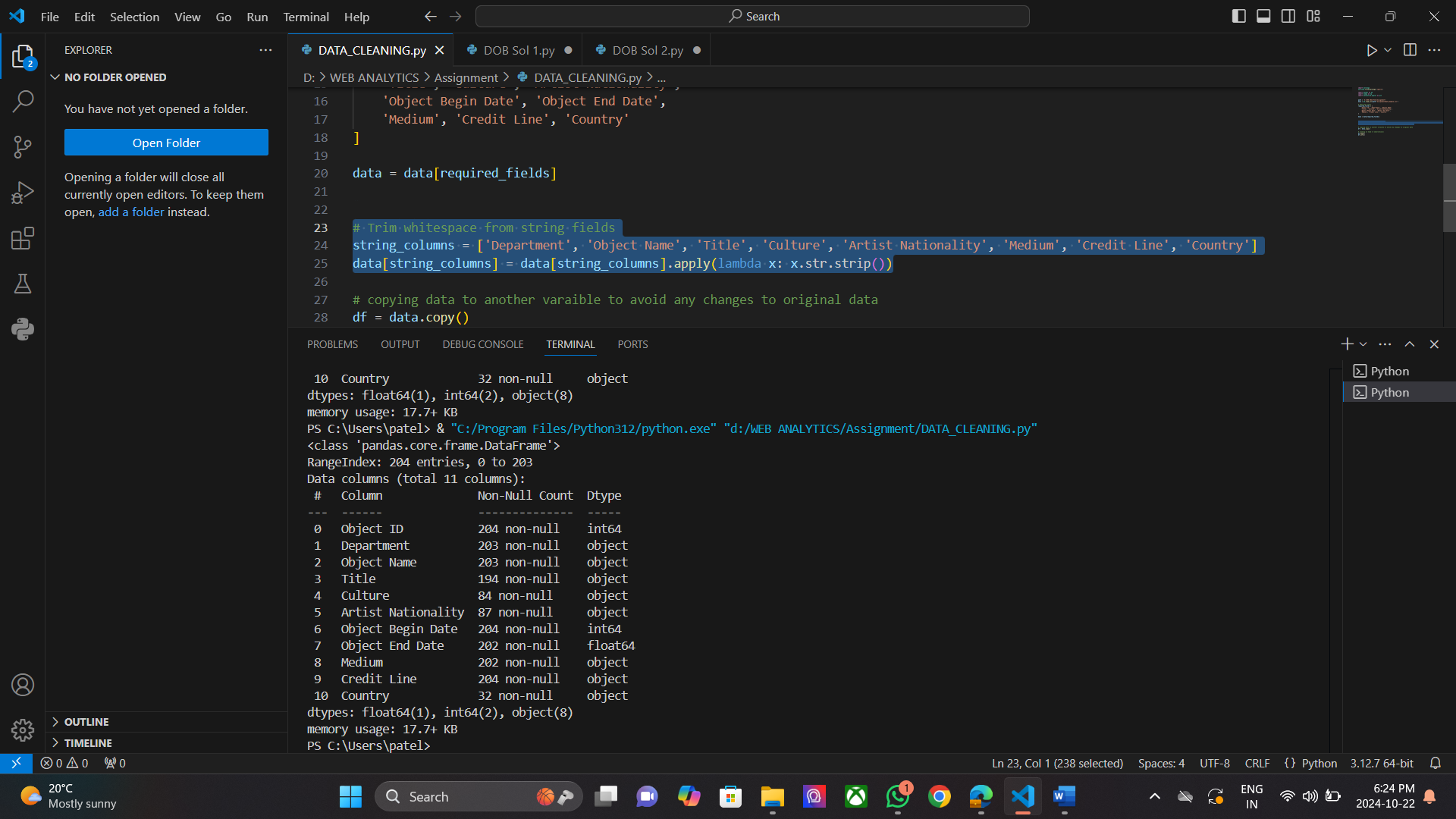
**2. trim record:**

Remove white space from the record, remove leading and trilling empty space and white space.

# Trim whitespace from string fields

string\_columns = ['Department', 'Object Name', 'Title', 'Culture', 'Artist Nationality', 'Medium', 'Credit Line', 'Country']

data[string\_columns] = data[string\_columns].apply(lambda x: x.str.strip())

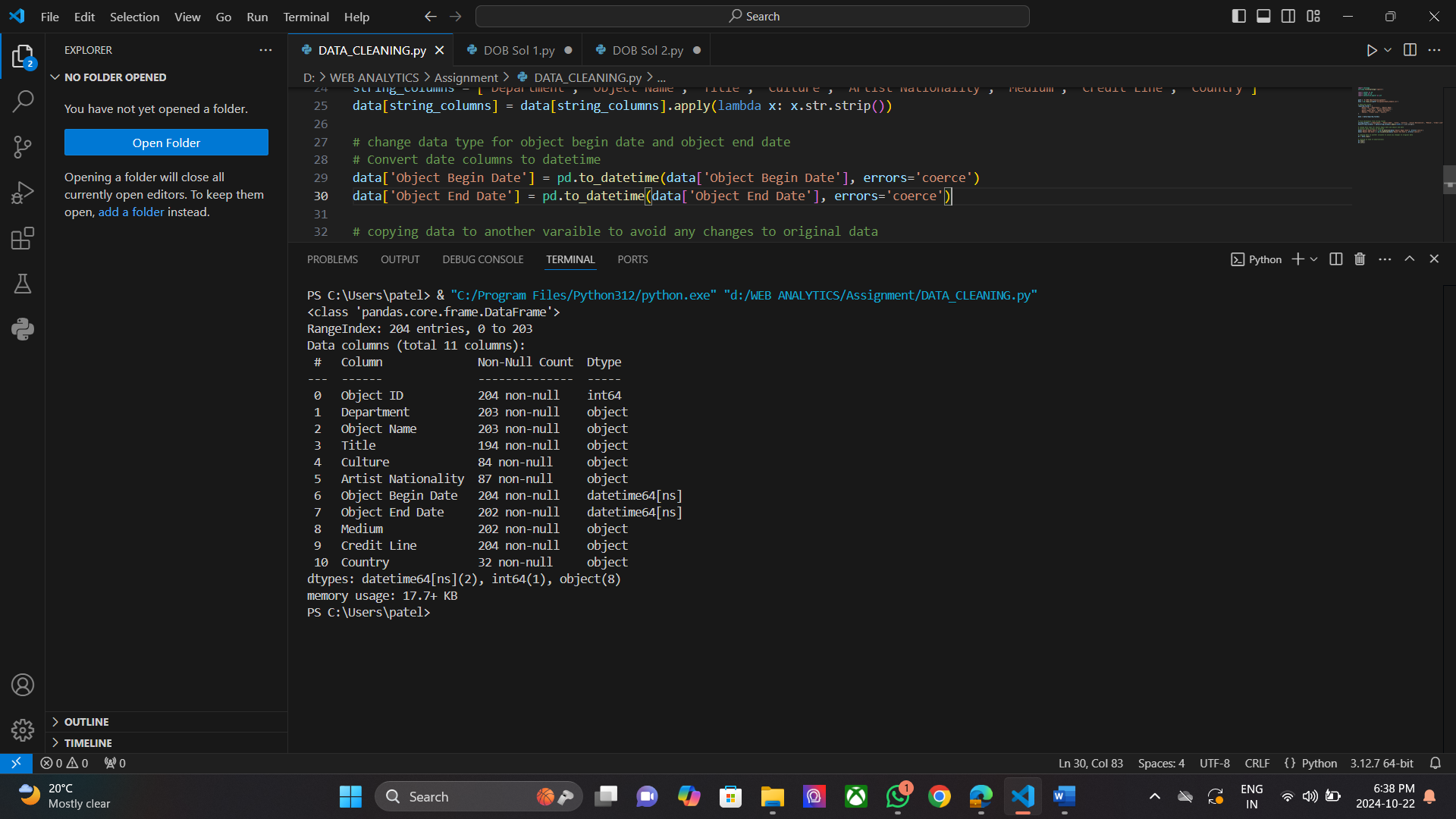


**3. change column datatype to date datatype for Object Begin date and Object End Date:**

Previously date column datatype id int and float so change that data type to date data type.

data['Object Begin Date'] = pd.to\_datetime(data['Object Begin Date'], errors='coerce')

data['Object End Date'] = pd.to\_datetime(data['Object End Date'], errors='coerce')



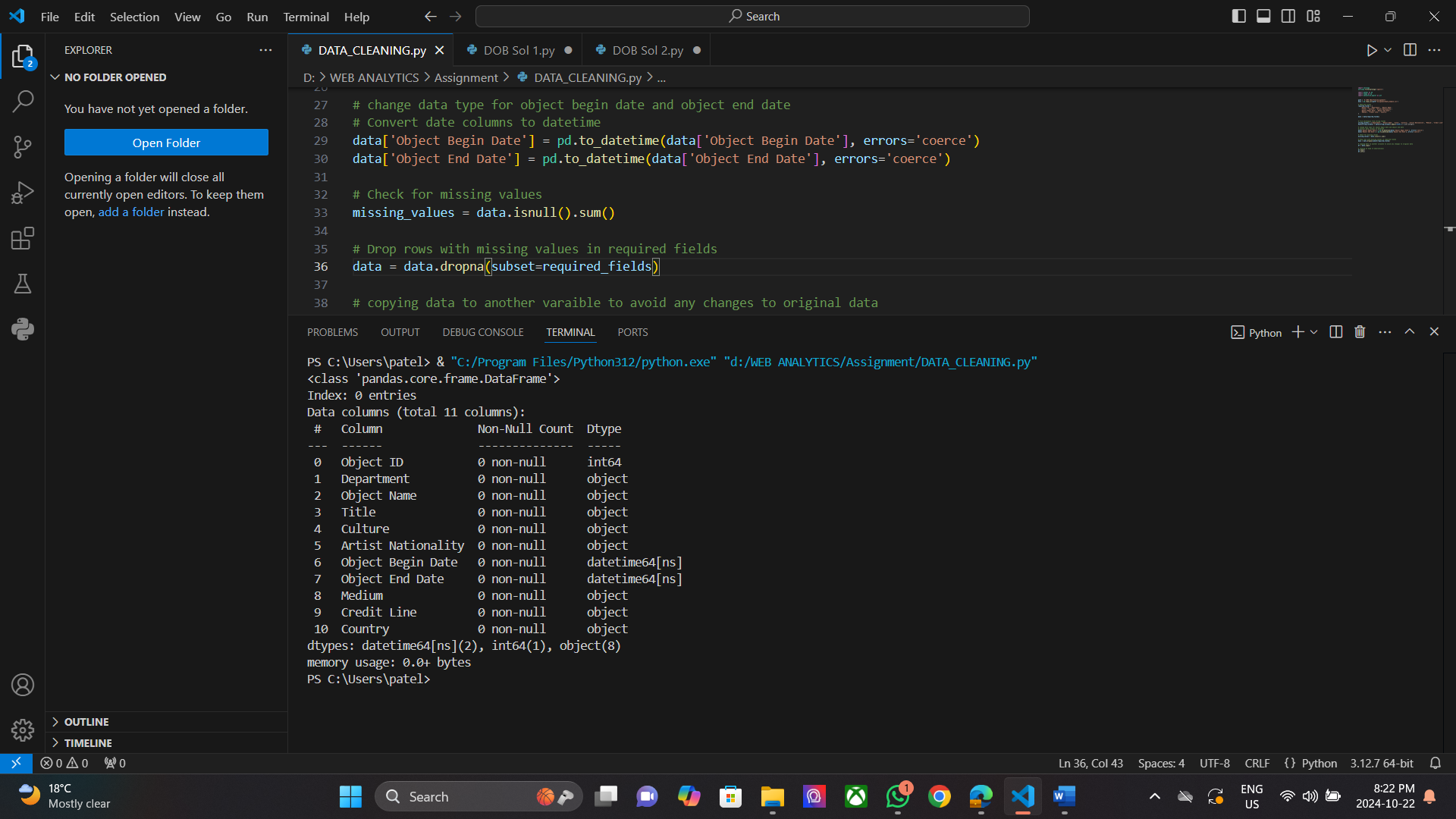
**4. remove all record with the missing value in required column**

# Check for missing values

missing\_values = data.isnull().sum()

# Drop rows with missing values in required fields

data = data.dropna(subset=required\_fields)



**5. Remove inconsistent record**

# Remove records where Object Begin Date is later than Object End Date

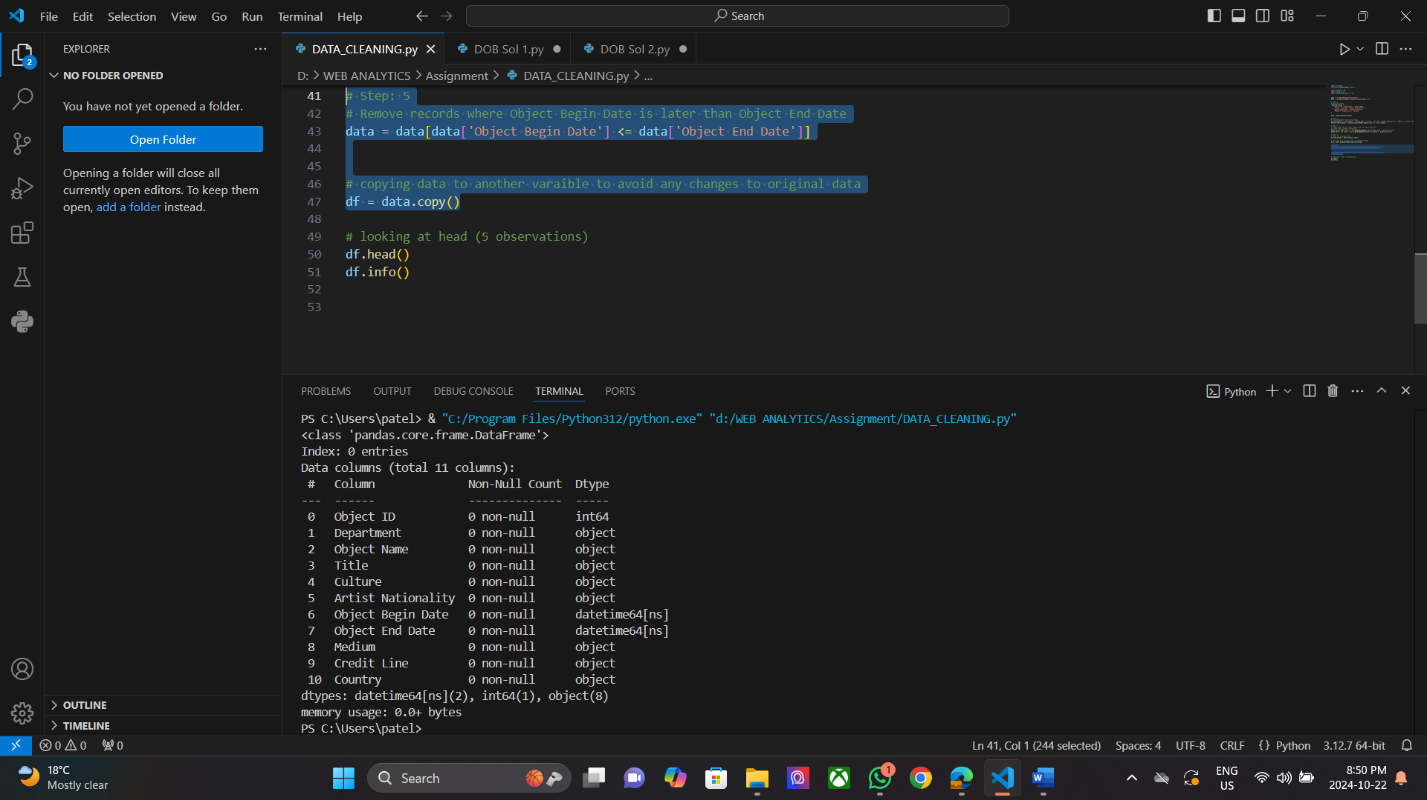
# Step: 5

# Remove records where Object Begin Date is later than Object End Date

data = data[data['Object Begin Date'] <= data['Object End Date']]

# copying data to another varaible to avoid any changes to original data

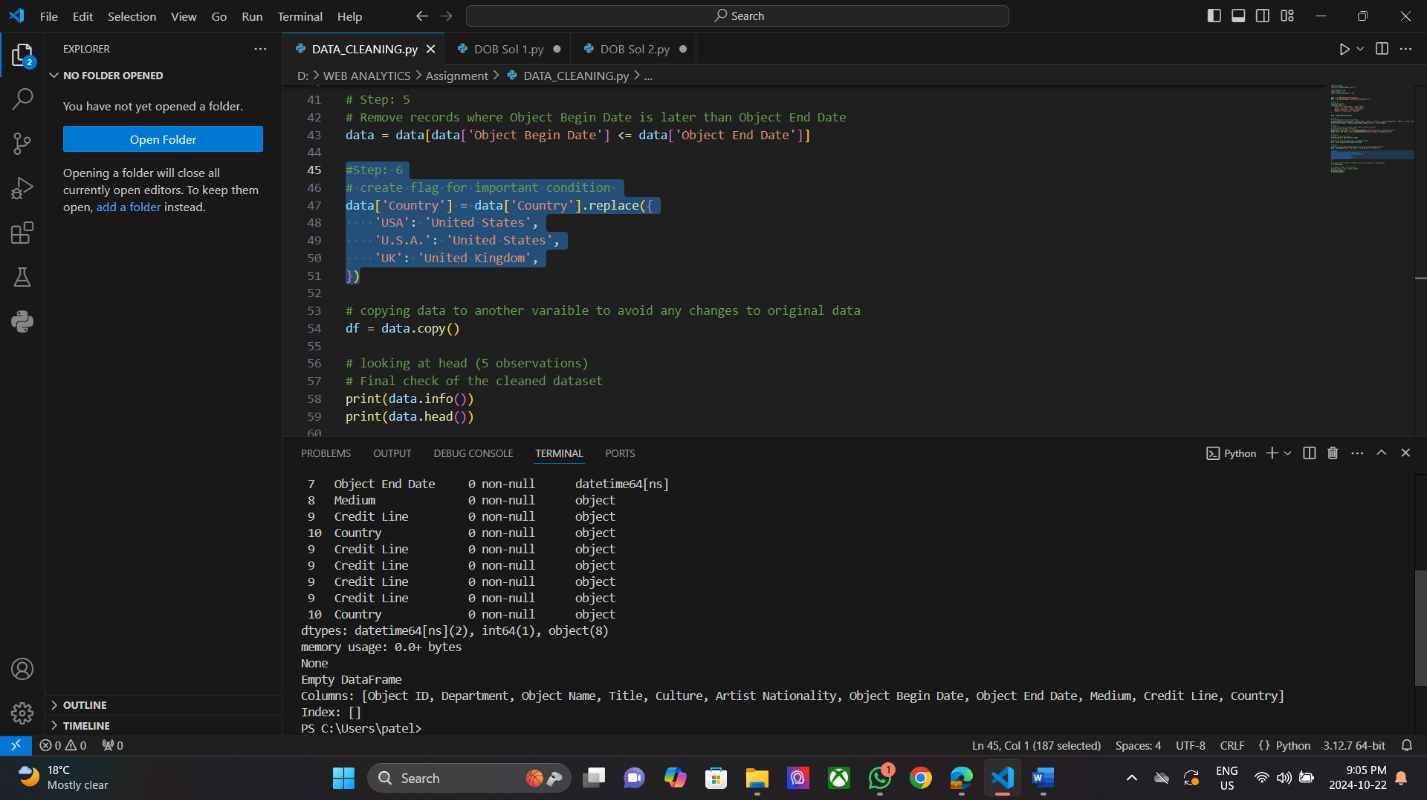
df = data.copy()



**6. Replace name**

Replacing short name with the full name

data['Country'] = data['Country'].replace({ 'USA': 'United States', 'U.S.A.': 'United States', 'UK': 'United Kingdom', })



**Task-2**

**1. What are the 3 Vs of Data and explain each one in detail?**

The 3 Vs of Data refer to Volume, Velocity, and Variety, which describe the key characteristics that define data. Volume relates to the massive amounts of data generated every second, Velocity refers to the speed at which this data is created and processed, while Variety highlights the diverse types of data that come from various sources. Understanding these three dimensions is crucial in the historical context of data visualization as they influence how data is collected, stored, and visualized over time.

**Volume**: Refers to the amount of data generated every second. With the proliferation of digital technology, organizations are faced with vast amounts of data. For instance, social media platforms generate millions of posts, comments, and interactions daily, which require significant storage and processing capabilities.

For example, [Todo List](http://todoist.com/) (the to-do manager I use) has roughly 10 million active installs, according to Android Play. That's not counting all the installs on the Web and iOS. Each of those users has lists of items -- and all that data needs to be stored. [Todo List](http://todoist.com/) is certainly not Facebook scale, but they still store vastly more data than almost any application did even a decade ago.

**Velocity:** Indicates the speed at which data is generated and processed. This is especially relevant for real-time data analytics. For example, financial markets generate data at incredibly high speeds, necessitating immediate processing to make timely trading decisions.

For example, Let's say you're running a marketing campaign and you want to know how the folks "out there" are feeling about your brand right now. How would you do it? One way would be to license some Twitter data from Gnip ([acquired by Twitter](https://blog.twitter.com/2014/twitter-welcomes-gnip-to-the-flock)) to grab a constant stream of tweets, and subject them to sentiment analysis.

**Variety**: Describes the different types of data that are collected from various sources. This includes structured data (like databases), unstructured data (like emails and videos), and semi-structured data (like JSON or XML files). For example, a retail business may collect customer purchase data, social media interactions, and online reviews, each representing different varieties of data.

For example, email messages. A legal discovery process might require sifting through thousands to millions of email messages in a collection. Not one of those messages is going to be exactly like another. Each one will consist of a sender's email address, a destination, plus a time stamp. Each message will have human-written text and possibly attachments.

**2. List capabilities of Business Intelligence systems?**

! Here’s a brief explanation of the capabilities of Business Intelligence (BI) systems:

**1. Data Management**

Some BI tools offer limited source connectivity, while others are flexible, allowing you to build custom connectors.

The platform must be able to draw data from cloud sources, including [eCommerce](https://www.selecthub.com/c/ecommerce-platforms/), [POS (point-of-sale)](https://www.selecthub.com/resources/pos-system/), [invoicing](https://www.selecthub.com/enterprise-resource-planning/accounting/billing-and-invoicing-software/), [billing](https://www.selecthub.com/c/medical-billing-software/), [procurement](https://www.selecthub.com/c/procurement-software/), [supply chain](https://www.selecthub.com/c/supply-chain-management-software/) and [CRM](https://www.selecthub.com/customer-relationship-management/crm-tools/) ([customer relationship management)](https://www.selecthub.com/customer-relationship-management/crm-tools/) systems.

**2. Data Lineage Tracking**

Knowing where your data comes from assures you of its reliability. Data lineage is a significant part of data quality management. That’s one less thing to worry about when many downstream processes rely on this information.

BI tools do you one better — they record this information, including the transformations it undergoes during its lifecycle, so you know what you’re getting.

**3. Data Visualizations**

[Data visualizations](https://www.selecthub.com/business-intelligence/data-visualization/) are any day preferable to tables with millions of rows and columns.

Interactive graphics make professional-looking presentations. Trying to explain sales projections? There’s a visual for it. Want to capture customer insight with [gamification](https://spinify.com/blog/why-gamification-is-important-its-benefits/)? There are visuals and animations for that.

**4. Cloud Deployment**

Deployment is how you implement the software — on-premise, in the cloud or as a hybrid installation.

The last one is an excellent solution if legacy systems weigh you down. Keeping proprietary data on-premise while computing in the cloud gives you the best of both worlds.

**5. Customization and Integration**

Enriching proprietary insights with transactional systems completes the business picture but leaves money on the table if external systems aren’t connected to your in-house platforms.

Many vendors provide customization options by offering [software development kits (SDKs)](https://builtin.com/software-engineering-perspectives/sdk) with built-in [APIs](https://www.ibm.com/topics/api). Additionally, an open architecture allows you to integrate the system with other platforms.

**6. Reporting**

Generating and distributing reports is a primary BI benefit. Its modern avatar is slicker and faster — ad hoc reports are the bread and butter of departments across organizations.

But we can’t write off canned reports yet. Sales, inventory, procurement, [human resources](https://www.selecthub.com/category/hris/), financial and marketing reports are regular in many organizations.

BI tools offer both types of reports, though some vendors refer to dashboards as reports, so it’s better to make your requirements clear at the onset.

**7. Dashboard**

[Dashboard solutions](https://www.selecthub.com/business-intelligence/businesses-need-dashboard-software/) are BI tools providing decision support with unified data views. Big data insights are easy to understand when viewed on a single dashboard. You get two for the price of one — data visualization and BI workflows — in one neat package.

**8. Collaboration**

Emails that get buried in your inbox, paper reports that increase your company’s carbon footprint and other outdated memo systems need not slow you down.

BI software supports working smarter by reducing turnarounds and inviting collaboration within the dashboard or report. Permissions configuration enables data access to the concerned teams so they can work together on documents, tasks and within workspaces.

**9. Predictive Analytics and Modeling**

[Predictive analytics](https://www.selecthub.com/business-intelligence/five-strategies-driving-revenue-growth-predictive-analytics/) is a top priority for organizations across the board. Forecasting trends can help you navigate the unpredictable business landscape in the face of pandemics, natural disasters, wars and other socio-economic disruptions.

**10. Mobile BI**: Allows access to BI tools and reports on mobile devices, facilitating on-the-go decision-making.

**3. Different types of data with example for each type?**

Here are four types of data:

**1. Nominal data**

This type of qualitative data is descriptive. Accordingly, nominal data features information that doesn't have quantitative values. Unlike numbers, you can't compare this type of data or order them. Instead, it's common to group them into distinct categories. Here are some examples of nominal data:

* **eye colour:** brown, black or blue
* **hair colour:** black, brown, grey or blonde
* **marital status:** single, married, divorced or widowed

**2. Ordinal data**

Ordinal data is another type of qualitative data. Unlike nominal data, the values in this classification have a natural set order or scale. This means that some sets might have numerical values and that you can order them because they've got a relative position to one another. Here are some common examples of this type of data:

* surveys that ask you to rate an experience on a scale of one to 10
* letter grades for examination results, such as A, B, C and D
* economic status, such as high, middle or low

**3. Discrete data**

This is a type of quantitative data, which is information that has numerical values. Discrete data generally only consists of fixed values, such as integers and whole numbers. Decimals can occur in this category, for instance, if you're collecting data on the profits that a company earns. Impossible or unrealistic values, such as a shoe size of 5.72, don't constitute discrete data. Here are some examples of discrete data:

* the total number of students in a class
* the cost of a mobile phone
* the total number of employees in an organisation

**4. Continuous data**

Continuous data is the opposite of discrete data, featuring sequences that have the potential to go on infinitely. You can divide this type of data into smaller pieces, and fractional numbers are common in this category. Here are some examples of continuous data:

* the height of a person
* the weight of a ship
* the temperature of a room
* the speed of a vehicle
* the time taken to complete a task

**4. Define data visualization.**

Data visualization translates complex data sets into visual formats that are easier for the human brain to comprehend. This can include a variety of visual tools such as:

* **Charts**: Bar charts, line charts, pie charts, etc.
* **Graphs**: Scatter plots, histograms, etc.
* **Maps**: Geographic maps, heat maps, etc.
* **Dashboards**: Interactive platforms that combine multiple visualizations.

The primary goal of data visualization is to make data more accessible and easier to interpret, allowing users to identify patterns, trends, and outliers quickly. This is particularly important in the context of big data, where the sheer volume of information can be overwhelming without

**5. What is a KPI and provide an example?**

A Key Performance Indicator (KPI) is a measurable target that indicates how individuals or businesses are performing in terms of meeting their goals. Reviewing and evaluating KPIs helps organizations determine whether or not they are on track for hitting their desired objectives.

By looking at several key indicators, which may include categories such as profits, sales numbers, employee turnover and average annual expenses, businesses can identify successes, as well as what is not working. Analyzing KPIs on a regular basis provides a solid overview of how well a business is performing, which allows the folks in charge to decide if current operations should be continued, or if a change of strategy is needed.

**Example of Marketing KPIs:**

* Monthly website traffic
* Page likes and comments
* Social media engagement rates
* Number of new monthly leads
* Click-through rate percentage

**Example of Customer Service KPIs:**

* Customer satisfaction score
* Customer retention rate
* Monthly support ticket submissions
* Average resolution time
* Cost per resolution

**6. What is a BI system?**

Business intelligence (BI) is a set of technological processes for collecting, managing and analyzing organizational data to yield insights that inform business strategies and operations.

**7. What are the 5 C's of Data for data preparation and the purpose of each?**

Five framing guidelines help us think about building data products. We call them the five Cs: consent, clarity, consistency, control (and transparency), and consequences (and harm). They’re a framework for implementing the golden rule for data.

**Consent**

You can’t establish trust between the people who are providing data and the people who are using it without agreement about what data is being collected and how that data will be used. Agreement starts with obtaining consent to collect and use data. Unfortunately, the agreements between a service’s users (people whose data is collected) and the service itself (which uses the data in many ways) are binary (meaning that you either accept or decline) and lack clarity. In business, when contracts are being negotiated between two parties, there are multiple iterations (redlines) before the contract is settled. But when a user is agreeing to a contract with a data service, you either accept the terms or you don’t get access. It’s non-negotiable.

There are many, many more examples of non-consensual data collection.

Google collected data from cameras mounted on cars to develop new mapping products. [AT&T and Comcast](https://www.techdirt.com/articles/20160609/12091134667/consumer-groups-say-att-comcast-violate-privacy-law-hoovering-up-cable-box-data-without-full-user-consent.shtml) both used cable set top boxes to collect data about their users, and Samsung collected voice recordings from TVs that respond to voice commands.

**Clarity**

Clarity is closely related to consent. You can’t really consent to anything unless you’re told clearly what you’re consenting to. Users must have clarity about what data they are providing, what is going to be done with the data, and any downstream consequences of how their data is used. All too often, explanations of what data is collected or being sold are buried in lengthy legal documents that are rarely read carefully, if at all. Observant readers of Eventbrite’s user agreement recently discovered that listing an event gave the company the right to send a video team, and exclusive copyright to the recordings. And the only way to opt out was by writing to the company. The backlash was swift once people realized the potential impact, and Eventbrite removed the language.

Most Twitter users know that their public tweets are, in fact, public; but many don’t understand that their tweets can be collected and used for research, or even that they are for sale.

**Consistency and trust**

Trust requires consistency over time. You can’t trust someone who is unpredictable. They may have the best intentions, but they may not honor those intentions when you need them to. Or they may interpret their intentions in a strange and unpredictable way. And once broken, rebuilding trust may take a long time. Restoring trust requires a prolonged period of consistent behavior.

Consistency, and therefore trust, can be broken either explicitly or implicitly. An organization that exposes user data can do so intentionally or unintentionally. In the past years, we’ve seen many security incidents in which customer data was stolen Yahoo!, Target, Anthem, local hospitals, government data, and data brokers like Experian, the list grows longer each day. Failing to safeguard customer data breaks trust—and safeguarding data means nothing if not consistency over time.

**Control and transparency**

Once you have given your data to a service, you must be able to understand what is happening to your data. Can you control how the service uses your data? For example, Facebook asks for (but doesn’t require) your political views, religious views, and gender preference.

All too often, users have no effective control over how their data is used. They are given all-or-nothing choices, or a convoluted set of options that make controlling access overwhelming and confusing. It’s often impossible to reduce the amount of data collected, or to have data deleted later.

**Consequences**

Data products are designed to add value for a particular user or system. As these products increase in sophistication, and have broader societal implications, it is essential to ask whether the data that is being collected could cause harm to an individual or a group. We continue to hear about unforeseen consequences and the “unknown unknowns” about using data and combining data sets. Risks can never be eliminated completely. However, many unforeseen consequences and unknown unknowns could be foreseen and known, if only people had tried. All too often, unknown unknowns are unknown because we don’t want to know.

**8. What are some Key Success Factors of a Successful BI Program and explain each factor?**

The following key steps ensure a smooth and successful BI implementation, from assessing your business requirements to deploying and maintaining your BI solution**.**

**1) Define Needs and Goals**

Business intelligence implementation starts with clearly understanding the organization’s overarching business objectives, goals, risks, and expectations and how implementing BI can contribute to achieving them. The first step towards implementation is to assess your current data infrastructure, including:

* Data sources
* Reporting practices
* Existing data silos or limitations.

Work closely with department heads and stakeholders to clearly define what you want to achieve with BI and where you would use BI insights. This could involve improving operational efficiency, gaining customer insights, or boosting sales. You can identify your specific BI needs through various techniques such as interviews, workshops, questionnaires, and other methods.

**2) Gather Requirements**

The next  BI implementation steps involve defining functional and non-functional requirements for the BI solution from all the stakeholders. The requirements can be further classified into mandatory and optional features. BI consultants play a critical role in this stage by interviewing stakeholders to collect and prioritize their needs, goals, and vision for a successful BI project implementation.

**3) Choose the Right BI Tools and Technologies**

Now that you have a precise understanding of your BI needs, the next step requires you to evaluate various BI tools and technologies based on factors such as:

* **Scalability**– Ensure the tools can accommodate your current data volume and anticipated future growth.
* **Functionality**– Prioritize tools that offer features relevant to your organization’s specific needs (e.g., data visualization, reporting, predictive analytics).
* **Integration**– Choose a tool that seamlessly integrates with your existing systems and data sources.
* **User-friendliness** – Choose tools with an intuitive interface for users with varying technical skills.

Identify necessary components to build a scalable BI ecosystem, including hardware resources, data storage solutions, and analytical tools to cater to your organization’s evolving requirements.

**4) Develop a BI Strategy and Plan**

Before diving into the technical aspects, ensuring that your business intelligence implementation strategy aligns with your organization’s overall business objectives, priorities, and strategic direction is essential. It is equally important to convince stakeholders of the value proposition of implementing BI with a robust strategy plan. Without a clear plan, employees may use inaccurate data, get conflicting instructions, and make assumptions about what things mean.

You must, therefore, create a detailed roadmap to achieve exceptional outcomes, including timelines, milestones, resource allocation, and other processes. While making the strategy, answer these three questions in detail,

* What is our current situation?
* What is our objective?
* What resources do we require?

The business intelligence implementation plan is a blueprint that provides guidelines on what would be needed to attain the goal, including scope, time frame, cost, and resources, among others. Project planning involves defining the work scope, identifying the project stakeholders, and establishing clear communication channels for efficient collaboration.

**5) Identify KPIs and Metrics**

Identify key measurable performance indicators to track the efficiency and status of workflows. Monitoring popular KPIs such as user adoption rate, Data quality score, and report usage metrics helps your organization track progress toward its goals and objectives. It also serves as a foundation for data-driven decision-making.

**6) Implement and Customize the BI Solution**

The step involves developing and tailoring the BI solution. It includes selecting the appropriate BI tools and configuring platforms for the organization’s specific data sources and infrastructure. This includes data preparation, setting up data storage, data integration, data management, customizing the BI platform, creating UI and dashboards for different departments, and performing quality assurance and optimization. The stage also involves integrating the BI solution into existing systems by automating the process with the help of data integration tools and establishing methods for extracting, transforming, and loading (ETL) data from various sources into the BI solutions.

**7) Run a Pilot Project**

Before deploying the BI system company-wide, testing it on a small scale, such as for one department or focus group, is recommended. This helps identify and mitigate risks, refine the implementation strategy, and help demonstrate the feasibility, functionality, and potential benefits of the BI solution once implemented organization wide.

**8) Gather Stakeholders’ Feedback**

You should initiate regular meetings or discussions with BI stakeholders to evaluate the current progress of the project and propose timely changes according to the suggestions. You can also launch a trial version to collect collective feedback from the end users and convert bottlenecks into opportunities for a successful BI strategy.

**9) Deploy and Launch the BI System**

The final step is to launch and deploy the BI system company-wide. This involves deploying the BI solution in several iterations, monitoring performance, and troubleshooting any issues that arise. The BI solution should be scalable to meet the organization’s growth needs.

**10) Monitor, Evaluate, and Continuously Improve**

Business intelligence implementation is an ongoing journey, not a one-time event. Your organization must be prepared to adapt and optimize your BI solution based on the following:

* **Evolving business needs:**BI implementation must aim to provide actionable insight to meet your growing business needs. As your business evolves, update your BI system to ensure it delivers the correct information and supports the organization’s strategic goals and data requirements.
* **Technological advancements:**Stay up to date on emerging BI technologies and incorporate innovations; organizations can adopt innovative technologies to enhance the system’s capabilities. This may include integrating ML algorithms for predictive analytics or leveraging cloud-based platforms for scalability and flexibility.
* **User feedback:**Continuously refine the system to meet changing user needs and preferences based on user input and feedback to ensure maximum value. This allows an organization to ensure that it remains user-friendly.

**Reference:**

1. *3 Vs of Big Data - (Data Visualization for Business) - Vocab, Definition, Explanations | Fiveable*. (2024). Fiveable.me[. https://library.fiveable.me/key-terms/data-visualization-for-business/3-vs-of-big-data](.%20https:/library.fiveable.me/key-terms/data-visualization-for-business/3-vs-of-big-data)
2. ‌*Volume, velocity, and variety: Understanding the three V’s of big data*. (2020, April 7). DataSource.ai. <https://www.datasource.ai/en/data-science-articles/volume-velocity-and-variety-understanding-the-three-v-s-of-big-data>
3. ‌Adair, B. (2021, February 11). *Business Intelligence Capabilities | BI Capabilities List For 2021*[. https://www.selecthub.com/business-intelligence/list-bi-capabilities/](.%20https:/www.selecthub.com/business-intelligence/list-bi-capabilities/)
4. 4 Types of Data (With Examples, Differences and Techniques). (n.d.). *Indeed Career Guide*. <https://sg.indeed.com/career-advice/career-development/types-of-data>
5. Jagdeesh. (2023, September 15). *Types of Data in Statistics - A Comprehensive Guide*. Machine Learning Plus. <https://www.machinelearningplus.com/statistics/types-of-data-in-statistics/>
6. ‌Geeksforgeeks. (2021, July 1). *What is Data Visualization and Why is It Important?* GeeksforGeeks. <https://www.geeksforgeeks.org/data-visualization-and-its-importance/>
7. ‌ Hennigan, L. (2024, June 15). *What Is A KPI? Definition & Examples*. Forbes. <https://www.forbes.com/advisor/business/what-is-a-kpi-definition-examples/>
8. ‌ Patil, D., Mason, H., & Loukides, M. (2018, July 24). *The five Cs*. O’Reilly Media. <https://www.oreilly.com/radar/the-five-cs/>
9. ‌ Admin. (2024, June 7). *10 Steps for Successful Business Intelligence Implementation*. Rishabh Software. <https://www.rishabhsoft.com/blog/business-intelligence-implementation>