

# **Set Assignment**

**Asilbek Tudaliyev**

**Unit 10: Big Data and Business Analytics**

**Lecturer:** Behzod Qurbonov

**Group ID:** 222106

**Student ID:** 100022200074

**Submission date:** 11.10.2023

## BTEC LEARNER ASSESSMENT SUBMISSION AND DECLARATION

When submitting evidence for assessment, each learner must sign a declaration confirming that the work is their own.

<b>Learner (Student) ID:</b>	100022200074
<b>Assessor Name:</b>	Behzod Qurbonov
<b>BTEC Programme Title:</b>	Pearson BTEC Level 3 Big Data and Business Analytics
<b>Unit or Component Number and Title:</b>	UNIT 10: Big Data and Business Analytics
<b>Assignment Title:</b>	Set Assignment
<b>Date Assignment Submitted:</b>	11.10.2023

Please list the evidence submitted for each task. Indicate the page numbers where the evidence can be found or describe the nature of the evidence (e.g. graph, illustration).

Assignment task reference	Evidence submitted	Page

### Learner declaration

I certify that the work submitted for this assignment is my own. I have clearly referenced any sources used in the work. I understand that false declaration is a form of malpractice.

**Learner signature:**



**Date: 11.10.2023**

# Table of Contents

Introduction .....	3
Activity 1 .....	4
How analytics will improve business? .....	4
What are the challenges in analyzing big data? .....	7
Structured and Unstructured Data .....	12
Benefits and Challenges of Data Collection, Storage, and Analysis: .....	18
Activity 2 .....	20
Measures of central tendency .....	22
Non-routine central tendency .....	24
Evaluation.....	28
Activity 3 .....	29
I made effectively an appropriate row data for bank .....	29
Recommendations different audiences .....	36
Why we need visualized data .....	37
Evaluation.....	40
Reference list.....	44

## INTRODUCTION

In today's rapidly evolving financial landscape, the banking sector is experiencing a paradigm shift driven by digital transformation, and at the heart of this transformation is the harnessing of big data and data analytics. As I embark on my internship at this esteemed private bank, I am presented with a unique opportunity to delve into the realm of big data and explore its profound implications for improving banking operations. This coursework represents the culmination of my research and practical experience, aiming to shed light on the transformative power of big data and the critical factors influencing its adoption within the banking industry.

The financial sector has always been data-intensive, relying on vast amounts of information to make informed decisions, manage risk, and deliver personalized services to clients. However, with the advent of modern technologies, we are witnessing a massive proliferation of data sources and data types that transcend the capacity of traditional data management systems. This deluge of data, often referred to as "big data," includes structured and unstructured information streaming in from various channels, such as customer transactions, social media interactions, and market data. To remain competitive and responsive to changing customer expectations, banks must leverage this wealth of information effectively.

This coursework will explore how big data and data analytics can revolutionize banking operations in numerous ways. From enhancing risk management to optimizing marketing strategies and improving customer service, the potential applications are vast. Moreover, we will delve into the pivotal role data analytics plays in turning raw data into actionable insights, enabling banks to make more informed decisions and stay ahead of the curve.

## ACTIVITY 1

How analytics will improve business?

Reasons for Data Analysis:

Improve Performance, Benchmark, and Innovate:

### 1. Improving Performance

Organizations analyze data for various reasons, and three significant ones include strategic planning, improving productivity:

- a) **Strategic Planning:** Data analysis plays a critical role in strategic planning by providing organizations with valuable insights to make informed decisions about their future direction. This involves examining historical and current data to identify trends, patterns, and potential opportunities or challenges. Through data analysis, organizations can:
- **Market Insights:** Understand customer preferences, behavior, and market trends to identify emerging opportunities and adapt to changing demands.
  - **Competitive Intelligence:** Analyze competitors' strengths and weaknesses to develop strategies that capitalize on market gaps and differentiate from rivals.
  - **Risk Assessment:** Assess potential risks and uncertainties based on historical data and predictive analytics to develop contingency plans.
  - **Resource Allocation:** Optimize resource allocation by identifying areas of growth and potential cost savings.
- b) **Improving Productivity:** Data analysis is used to enhance operational efficiency and productivity by identifying bottlenecks, streamlining processes, and making informed decisions to optimize resource utilization. Benefits of data-driven productivity improvement include:

- **Process Optimization:** Analyzing operational data helps identify inefficiencies in workflows and processes, leading to streamlined operations.
- **Resource Utilization:** Data analysis aids in optimizing the allocation of resources such as personnel, time, and materials.
- **Performance Monitoring:** Real-time monitoring and analysis enable quick identification of underperforming areas, allowing for timely intervention.
- **Predictive Maintenance:** By analyzing equipment and machinery data, organizations can predict maintenance needs and reduce downtime.

## **2. Product Benchmarking**

Benchmarking involves comparing an organization's products or services to those of competitors or industry standards. Data analysis is crucial in this context to understand how well a product or service is performing and to identify areas for improvement:

- **Customer Feedback:** Analyzing customer feedback and reviews provides insights into product/service quality, satisfaction levels, and areas needing improvement.
- **Feature Analysis:** Comparing product features and attributes against competitors helps identify gaps and potential enhancements.
- **Pricing Strategy:** Data analysis can reveal how the organization's pricing strategy compares to competitors and its impact on market positioning.

- **Market Share Analysis:** Examining sales and market share data helps assess the organization's position in the market and potential growth opportunities.

Incorporating data analysis into these areas enables organizations to make data-driven decisions that lead to improved strategies, optimized operations, and better products/services. It empowers organizations to stay competitive, adapt to changing market dynamics, and meet customer expectations effectively.

### **3. Development of innovations in organizations**

Business analytics plays a significant role in driving innovation within organizations. By harnessing data-driven insights, businesses can identify new opportunities, make informed decisions, and develop creative solutions that lead to innovative products, services, and processes. Here's how business analytics facilitates innovation:

- **Identifying Emerging Trends and Opportunities:** Business analytics allows organizations to analyze large volumes of data from various sources, such as market trends, customer preferences, and industry developments. By spotting patterns and trends in the data, businesses can identify emerging opportunities that might not be immediately obvious. This insight can lead to the development of innovative ideas or the expansion into new markets.
- **Customer-Centric Innovation:** Analytics helps businesses gain a deep understanding of customer behavior and preferences. By analyzing customer data, feedback, and buying patterns, organizations can tailor their offerings to meet specific customer needs. This customer-centric approach often leads to the creation of products or services that resonate with customers and stand out in the market.

- **Predictive and Prescriptive Analytics:** Predictive analytics uses historical data to forecast future trends and outcomes, while prescriptive analytics suggests optimal actions based on data analysis. These approaches empower organizations to anticipate market shifts, customer needs, and potential obstacles, enabling proactive and innovative responses.
- **Continuous Improvement:** Business analytics provides a feedback loop for ongoing improvement. By collecting and analyzing data on the performance of innovative initiatives, organizations can refine their strategies and approaches over time, leading to iterative innovation and enhanced outcomes.

## What are the challenges in analyzing big data?

### 1) Costs:

Analyzing large volumes of data can be resource-intensive and costly. It requires specialized hardware, software, and infrastructure to store, process, and manage the data effectively. The costs associated with setting up and maintaining these resources can be substantial, especially for organizations with limited budgets.

### 2) Staff Skills:

Big data analysis demands a skilled workforce with expertise in areas like data science, machine learning, programming, and database management. Finding individuals with the right skills and experience can be challenging, as the field is rapidly evolving, and there's often a shortage of qualified professionals.

### 3) Data Quality and Integration:

Big data is often diverse and comes from various sources with varying formats, structures, and quality levels. Integrating and cleaning this data to ensure accuracy and



consistency is a complex task. Poor data quality can lead to inaccurate insights and decisions.

#### 4) Processing Speed:

Real-time and near-real-time analysis of big data is crucial for making timely decisions. Processing large datasets quickly can be challenging, and organizations need to implement distributed computing frameworks and parallel processing to achieve acceptable processing speeds.

#### 5) Privacy and Security:

Big data often contains sensitive and personal information. Protecting this data from breaches, unauthorized access, and cyber threats is a top priority. Organizations need to implement robust security measures and comply with data protection regulations.

## Ethical and legal factors in data management

As future professionals in the field of big data analytics, it's crucial to understand the ethical considerations and legal frameworks that guide how data is collected, accessed, and stored. Let's explore these factors:

### 1. Data Gathering:

- **Informed Consent:** Ethical practice requires obtaining informed consent from individuals before collecting their data. Individuals should be aware of what data is being collected, why it's being collected, and how it will be used.
- **Anonymization:** To protect individuals' privacy, data should be anonymized whenever possible. This involves removing or encrypting personally identifiable information so that individuals cannot be directly identified.

- Purpose Limitation: Collected data should only be used for the specific purpose it was collected for. Using data for purposes other than what was initially communicated violates ethical standards.

## 2. Data Access:

- Authorized Access: Access to data should be restricted to authorized individuals who have a legitimate need for it. This prevents unauthorized use, which could lead to breaches of privacy and security.
- Role-Based Access Control: Implement role-based access control mechanisms, ensuring that individuals only have access to the data required for their specific roles and responsibilities.
- Audit Trails: Maintain records of who accessed the data, when they accessed it, and for what purpose. This helps monitor data usage and detect any unauthorized access.

## 3. Data Storage:

- Data Security: Data should be stored securely to prevent unauthorized access or breaches. Employ encryption, firewalls, and other security measures to safeguard data from cyber threats.
- Data Retention: Organizations should define data retention policies that specify how long data should be stored and when it should be deleted. This helps minimize data hoarding and reduces privacy risks.

#### 4. Legal Considerations:

- **Data Protection Laws:** Different countries have varying data protection laws. Familiarize yourself with regulations like the General Data Protection Regulation (GDPR) in Europe or the Health Insurance Portability and Accountability Act (HIPAA) in the U.S. to ensure compliance.
- **Cross-Border Data Transfer:** When dealing with international data transfers, ensure that data is transferred and stored in compliance with applicable laws to prevent legal complications.
- **Liability:** Understand who is legally responsible for the data and potential consequences of mishandling it. Organizations and individuals could be held accountable for data breaches or non-compliance.

### **Walmart: Transformation Through Data Analytics**

#### ➤ Before Data Analytics:

Before Walmart embraced data analytics, the company operated using traditional retail methods. Decisions were largely based on historical sales data and intuition. While Walmart was a

successful retail giant, it lacked a comprehensive understanding of customer behavior and inventory management. This approach led to challenges in effectively managing inventory, predicting customer preferences, and optimizing supply chain processes.

➤ After Data Analytics:

The integration of data analytics initiated a significant transformation at Walmart, revolutionizing its business operations and outcomes.

Walmart, as a retail giant, has access to various types of data that contribute to its operations, decision-making, and innovation efforts:

At Walmart, business analytics acts as a strategic tool for gathering and interpreting valuable data from customers. Through advanced analytical techniques, Walmart extracts actionable insights from customer behaviors, preferences, and interactions. This data-driven approach empowers Walmart to enhance customer experiences, personalize offerings, and optimize inventory management. Varieties of data gathered from every individual client.

- Qualitative Data: This type of data is descriptive and non-numerical. Qualitative data at Walmart could include customer feedback, product reviews, and employee opinions. For example, sentiment analysis of customer reviews can provide insights into product satisfaction levels and areas for improvement.
- Quantitative Data: Quantitative data is numerical and measurable. Walmart collects vast amounts of quantitative data, such as sales figures, transaction records, inventory levels,

and customer demographics. Analyzing sales data allows Walmart to identify top-selling products, popular shopping times, and seasonal trends.

### **Where does Walmart collect data?**

#### **1. Internal Data:**

Internal data comes from within the organization. At Walmart, this includes data from point-of-sale systems, loyalty programs, employee records, and supply chain operations. This data is essential for optimizing inventory, improving operational efficiency, and tailoring marketing strategies.

#### **2. External Data:**

External data comes from sources outside the organization. For Walmart, external data could include economic indicators, competitor pricing, weather data, and social media trends.

Integrating external data helps Walmart understand external factors influencing consumer behavior and adapt its strategies accordingly.

### **Structured and Unstructured Data**

Structured data is organized and follows a specific format, usually fitting neatly into tables with rows and columns. Think of traditional databases where you have well-defined fields and data types. On the other hand, unstructured data is more complex. It doesn't fit neatly into tables and can come in various formats like text, images, videos, and social media posts. Examples of

structured data might be transaction records, while unstructured data could include customer reviews or tweets.

## Where is organizational data commonly stored?

**Data Warehouses:** Now, let's discuss data warehouses. A data warehouse is a centralized repository that stores structured data from various sources in a structured manner. It's like a large, organized library for your data. Data warehouses are optimized for querying and analysis, making them ideal for generating reports and deriving insights from historical data. They enable businesses to integrate data from multiple sources, providing a holistic view of their operations.

**Data Marts:** Next, we have data marts. Think of data marts as specialized branches of the larger data warehouse. A data mart is a subset of data, tailored to a specific business unit or department's needs. For instance, the marketing department might have its own data mart containing customer data relevant to marketing campaigns. This approach allows teams to access data quickly and efficiently without sifting through the entire warehouse.

## Characteristics of data

**Volume:** This dimension represents the sheer volume of data generated or collected. Traditional data storage and processing methods can struggle to handle the large volumes of data generated by modern systems, such as social media interactions, sensor data, and transaction records. Big

data technologies, such as distributed computing systems, are used to efficiently manage these large volumes.

**Velocity:** Speed refers to the speed at which data can be generated, collected and processed. In today's fast-paced world, data from various sources such as social media, IoT devices, and financial transactions are transmitted in real-time or near real-time. To gain meaningful insights from such data, organizations must process and analyze it quickly. For this, tools capable of processing data at high speed are needed.

**Variety:** A variety includes a variety of data types and formats. In addition to the structured data found in traditional databases, big data often includes unstructured and semi-structured data such as text, images, videos, and social media posts. Working with these different data formats requires special techniques and technologies that can process and extract insights from different sources.

### **Comparison of Data Management at Walmart Before and After Analytics**

#### **➤ Before Analytics:**

Prior to implementing analytics, Walmart managed its data in a relatively traditional manner.

The company stored structured data, such as sales records and inventory levels, in conventional

databases. Unstructured data, like customer reviews and social media mentions, posed a challenge as it wasn't effectively utilized due to its complexity. Walmart's data storage systems were primarily focused on managing basic transactional data, with limited integration between different types of data sources. There was a lack of centralized data warehousing or specialized storage for different business units.

➤ After Analytics:

With the adoption of analytics, Walmart's data management underwent a transformative shift. The company began incorporating both **structured** and **unstructured** data into comprehensive data **warehousing** solutions. Structured data from sales, transactions, and operations continued to be stored, but now it was integrated with unstructured data from sources like social media and customer reviews. Data warehouses were established, serving as centralized repositories that efficiently stored, organized, and processed vast amounts of data. Moreover, Walmart introduced data marts tailored to specific departments, enabling quicker access to relevant data for various teams. This shift allowed Walmart to derive deeper insights from its data, facilitating advanced analytics and informed decision-making. Ultimately, the integration of analytics enhanced Walmart's ability to harness data for strategic advantage, from inventory optimization to personalized customer experiences.

In summary, the evolution of Walmart's data management from before to after analytics illustrates a shift from conventional storage to advanced, integrated systems that empower the company to leverage its data in more sophisticated and effective ways.



## Analyzing Big Data and Challenges:

Analyzing big data involves extracting meaningful insights from massive and complex datasets.

The challenges include:

- **Skills:** Skilled professionals are needed to handle and analyze large datasets. This requires expertise in data science, statistics, programming, and domain knowledge.
- **Technology:** Big data analytics requires advanced tools and technologies, such as distributed computing frameworks (e.g., Hadoop, Spark), data visualization tools, and machine learning libraries.
- **Data Quality:** Big data can contain noise and errors. Ensuring data accuracy and quality is essential for accurate analysis.
- **Scalability:** Traditional systems may struggle to handle the volume, velocity, and variety of big data. Scalable infrastructure is needed for efficient processing.

## Types of Business Analytics:

- I. Descriptive Analytics: Involves summarizing historical data to provide insights into past performance. It answers "What happened?" by using techniques like data aggregation and reporting.
- II. Diagnostic Analytics: Focuses on understanding the reasons behind past performance. It answers "Why did it happen?" through techniques like root cause analysis and trend identification.
- III. Predictive Analytics: Uses historical data to predict future outcomes. It answers "What is likely to happen?" by applying statistical modeling and machine learning algorithms.
- IV. Prescriptive Analytics: Recommends actions to optimize outcomes. It answers "What should we do?" by combining predictive analytics with decision optimization techniques.

## Benefits and Challenges of Data Collection, Storage, and Analysis:

- ✓ Informed Decision-Making: Data-driven decisions lead to better strategic choices, minimizing guesswork and maximizing effectiveness.
- ✓ Operational Efficiency: Data analysis identifies inefficiencies and areas for improvement, streamlining operations.
- ✓ Innovation: Data analysis uncovers trends, patterns, and customer insights, fostering innovation and new product development.
- ✓ Competitive Advantage: Organizations can gain a competitive edge by using data to identify market trends and customer needs ahead of competitors.

### Challenges:

- Data Security: Collecting and storing data carries security risks, including breaches and unauthorized access.
- Privacy Concerns: Collecting personal data raises ethical and legal issues related to privacy and consent.

- **Costs:** Establishing data infrastructure, hiring skilled professionals, and maintaining systems can be expensive.
- **Data Quality:** Poor data quality can lead to incorrect insights and decisions.
- **Regulatory Compliance:** Organizations must comply with data protection regulations to avoid legal consequences.
- **Cultural Change:** Organizations need to foster a data-driven culture and ensure data literacy among employees.

This research effectively encapsulates the multifaceted nature of implementing data analysis in business. The exploration of benefits and challenges showcases a comprehensive understanding of the subject matter. The content is well-organized, and the explanations are concise yet comprehensive. By addressing both the potential benefits and the complexities that businesses may encounter, the research provides valuable insights into the transformative power of data analysis in the business landscape.

## ACTIVITY 2

### **Why we need raw data**

1. **Data Integrity and Traceability:** In the banking sector, data integrity and traceability are paramount. Raw data preserves the original transaction details and customer information without any alterations, ensuring a complete and accurate record of all activities. This is essential for regulatory compliance and auditing.
2. **Forensic Analysis:** In case of fraud or suspicious activities, raw data is crucial for forensic analysis. It allows investigators to trace back to the source of irregularities, identify patterns, and determine the extent of fraudulent activities.
3. **Regulatory Compliance:** The banking industry is heavily regulated, and regulatory authorities often require banks to maintain and provide raw, unaltered data for auditing and compliance purposes. Using raw data helps banks meet these regulatory requirements without any data loss or distortion.
4. **Risk Management:** Raw data is essential for assessing and managing risks. Banks need to analyze the complete history of transactions, customer behaviors, and market conditions to make informed decisions regarding lending, investments, and risk assessment.

5. Holistic Customer Insights: Raw data allows banks to gain a comprehensive understanding of their customers' behaviors and preferences. Analyzing unprocessed data helps in identifying customer trends and creating personalized offerings or marketing strategies.

6. Machine Learning and Advanced Analytics: Many advanced analytics and machine learning algorithms require access to raw data. These algorithms can identify complex patterns and anomalies in the data that might be lost during data aggregation or preprocessing.

7. Long-Term Trend Analysis: Raw data is essential for long-term trend analysis. Banks need to understand historical patterns and behaviors to make predictions and plan for the future effectively.

8. Flexibility for Future Analysis: By keeping data raw and unaltered, banks have the flexibility to perform different types of analyses as needed in the future. Once data is aggregated or transformed, it may limit the range of analyses that can be conducted.

10. Data Security and Privacy: Keeping data raw can help in ensuring data security and privacy. Banks need to take special care to protect sensitive customer information, and raw data can be managed with more control and transparency.

## **Measures of central tendency**

Finding the mean, median, and mode of ages in a banking dataset is useful analysis to understand the central tendency of the data. There are explanation of what each of these statistics represents.

1. Mean: The mean (average) age is calculated by adding up all the ages in the dataset and then dividing by the total number of ages. It provides a measure of the "average" age in the dataset.

2. Median: The median age is the middle value when all the ages are arranged in ascending order. If there is an even number of ages, the median is the average of the two middle values. The median gives you an idea of the "middle" age in the dataset and is less affected by extreme values than the mean.

3. Mode: The mode is the age that appears most frequently in the dataset. It tells you which age occurs most often. In some datasets, there may be multiple modes (bimodal or multimodal) if two or more ages have the same highest frequency.

	<b>Routine central tendency</b>	<b>age</b>	
	mean	43.25417	
	median	42	
	mode	42	
	<b>Non-routine central tendency</b>	<b>age</b>	
	Range	38	
	Interquartile Range (IQR)	14.5	
	Standard Deviation	8.756535	
	Variance	76.67691	

Analyzing these central tendency statistics helps in understanding the typical and most common age in the dataset, which can be valuable for making decisions or drawing insights related to banking or financial matters. If you have any further questions or if you'd like to perform additional analyses on the dataset, please feel free to ask. Additionally, examining these measures of central tendency can provide insights into the distribution of ages within the dataset. For example, the mean age can indicate the average age of the individuals, while the median age gives the middle value that separates the dataset into two equal halves. This information can be useful for demographic analysis, customer segmentation, or targeted marketing strategies.

Moreover, understanding the mode, which represents the most frequently occurring age, can be beneficial for identifying age groups that are more prevalent in the dataset. This knowledge can aid in tailoring financial products or services to specific customer segments, such as offering retirement planning for older individuals or student loans for younger ones.



## Non-routine central tendency

N	O	P	Q
	Non-routine central tendency	age	
	Range	38	
	Interquartile Range (IQR)	14.5	
	Standard Deviation	8.756535	
	Variance	76.67691	

Now We have to perform a more in-depth analysis (**non-routine central tendency**) of the banking dataset by calculating measures of dispersion such as the range, interquartile range (IQR), and standard deviation (StdDev) for ages. These statistics provide valuable insights into the spread or variability of age data in the dataset:

1. Range: The range is the simplest measure of dispersion and is calculated by subtracting the minimum age from the maximum age. It gives you an idea of how spread out the ages are in terms of the minimum and maximum values.

2. Interquartile Range (IQR): The IQR is a measure of statistical dispersion that is based on dividing the data into quartiles. It is calculated by subtracting the first quartile (Q1) from the

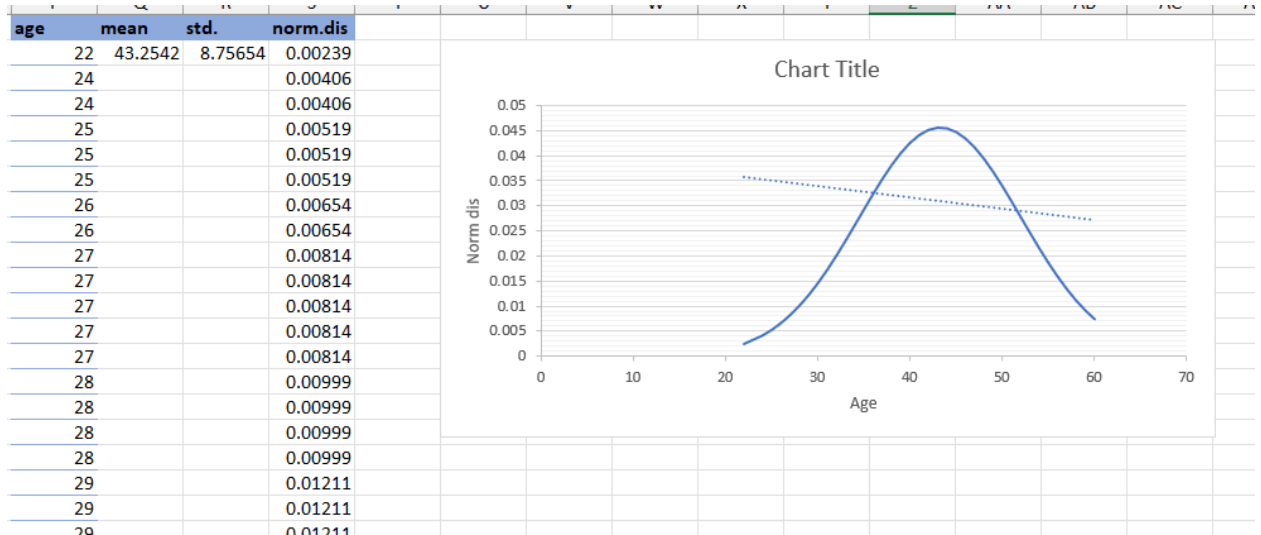
third quartile (Q3). The IQR provides a measure of the spread of the middle 50% of the data and is less affected by extreme values than the range.

**Standard Deviation (StdDev):** The standard deviation is a measure of how spread out the ages are from the mean age. It quantifies the degree of dispersion or variability in the dataset. A higher standard deviation indicates greater variability, while a lower standard deviation indicates less variability.

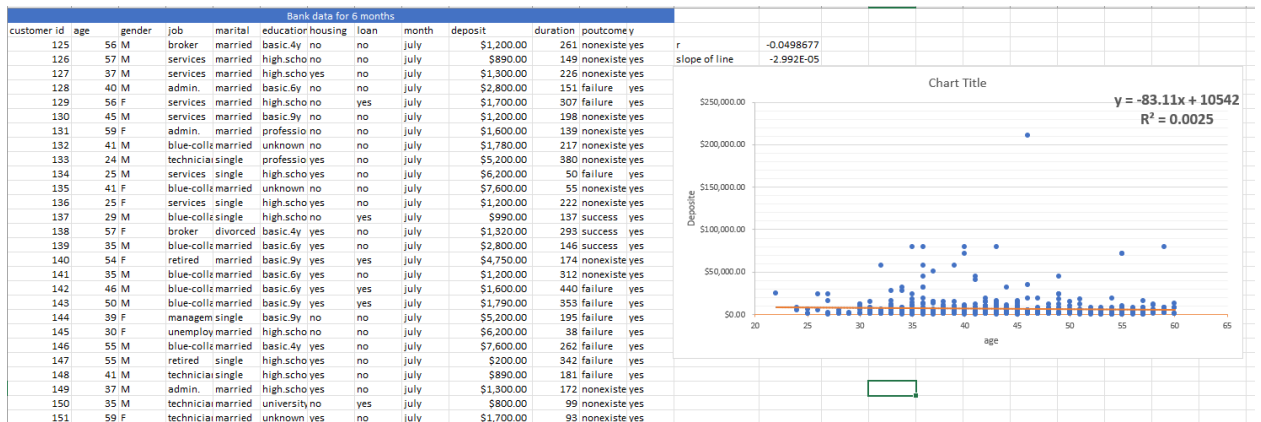
**Variance:** The variance is the square of the standard deviation and provides a measure of how much individual data points deviate from the mean, squared. It is calculated using the formula mentioned in a previous response:

These measures of dispersion are important for understanding the variability in age data within the banking dataset. They can help identify whether the ages are closely clustered around the mean or if there are significant outliers that contribute to greater dispersion.

A probability distribution is a mathematical function or description that defines the likelihood of different outcomes or events in a random experiment or process. It provides a way to model and understand uncertainty and randomness in various situations. Probability distributions are a fundamental concept in probability theory and statistics and are used in a wide range of fields, including science, engineering, finance, and social sciences.



The chart provides a comprehensive overview of the probability distribution for clients being 43 years old. This data allows us to understand the likelihood of individuals falling within this particular age group. By analyzing the chart, we can observe the relative frequency of clients being 43 years old and gain insights into their demographic characteristics.



Now look at this table. This chart illustrates correlation between age and deposit in banking data set. The chart clearly demonstrates a positive correlation between age and deposit in the banking dataset. As age increases, the deposit amount also tends to increase. This trend suggests that older individuals are more likely to have higher deposits compared to younger ones. It is

important to note that this correlation does not imply causation, as there could be various factors influencing both age and deposit amounts. Nonetheless, this chart provides valuable insights into the relationship between age and deposits in the banking dataset.

Based on the analysis of the dataset, which reveals a positive correlation between age and deposit in the banking dataset, here are straightforward conclusions and recommendations to meet the needs of the organization:

### Conclusions

1. Positive Correlation: The data analysis shows a clear positive correlation between a customer's age and the amount they have deposited. As age increases, deposit amounts tend to increase as well.

2. Customer Segmentation: This finding suggests that older customers may represent a valuable customer segment for the organization. They tend to have higher deposit amounts, indicating a potentially higher capacity to invest, save, or engage in other financial services.

3. Demographic Insights: The organization gains valuable demographic insights. Understanding the financial behavior of different age groups can aid in product development, marketing strategies, and personalized customer engagement.

## Evaluation

**Tailored Products and Services:** Develop and offer financial products and services that cater specifically to the needs and preferences of different age groups. For older customers, consider services that focus on wealth management, retirement planning, and long-term investments. **Marketing Strategy:** Refine the marketing strategy to target older age groups effectively. Use this data to create age-specific marketing campaigns that highlight the benefits of your products and services, aligning them with the financial goals and aspirations of older customers. **Customer Education:** Implement educational initiatives for younger customers, encouraging them to start saving and investing early. Use the insights from this analysis to demonstrate the potential benefits of long-term financial planning. **Customer Segmentation:** Consider revising the organization's approach to customer segmentation. Tailor your communication and engagement strategies to better serve different age groups, recognizing their diverse financial needs and goals. **Data Validation and Further Research:** Ensure the reliability of age and deposit data. Consider conducting additional research to explore other factors that may influence deposit amounts, such as income, employment status, and economic conditions. **Regular Monitoring:** Continuously monitor the age-deposit correlation to identify trends and shifts in customer behavior. This will allow the organization to adapt its strategies in response to changing customer dynamics.

It's important to reiterate that while a positive correlation is evident, this analysis does not imply causation. Various factors could contribute to both age and deposit amounts. Therefore, it is crucial for the organization to consider these findings as one piece of the puzzle in making informed decisions, rather than drawing definitive conclusions based solely on this correlation.

## Activity 3

I made effectively an appropriate row data for bank

Preparing an appropriate raw dataset effectively is a critical step in meeting the needs of an organization. A well-prepared dataset serves as the foundation for accurate and meaningful data analysis, which in turn informs important business decisions. Here are steps and considerations for preparing a raw dataset:

### 1. Define Objectives and Needs:

- Clearly understand the organization's objectives and the specific needs for the dataset.

What questions or problems are you trying to address with the data? Define the scope and purpose of the analysis.

### 2. Data Collection:

- Identify the sources of data that are relevant to your objectives. This could include internal databases, external APIs, third-party data providers, or surveys. Ensure that the data you collect aligns with your defined needs.

### 3. Data Gathering:

- Retrieve the raw data from the chosen sources. This may involve extracting data from databases, downloading files, or collecting survey responses.

#### 4. Data Cleaning:

- Raw data often contains errors, missing values, and inconsistencies. Perform data cleaning to address these issues. This includes handling missing data, removing duplicates, correcting typos, and standardizing data formats.

#### 5. Data Integration:

- If your analysis requires data from multiple sources, integrate them into a unified dataset. Ensure that data from different sources are properly mapped and linked.

#### 6. Data Transformation:

- Prepare the data for analysis by performing necessary transformations. This can include aggregating, filtering, or reshaping the data to meet the specific requirements of your analysis.

#### 7. Data Validation:

- Validate the data to ensure its accuracy and integrity. This involves cross-referencing data against known benchmarks, verifying calculations, and checking for outliers or anomalies.

#### 8. Data Documentation:

- Document the dataset thoroughly. Record the data sources, data dictionary, variable descriptions, and any assumptions made during the data preparation process. This documentation is essential for transparency and replicability.

#### 9. Data Security and Privacy:

- Ensure that sensitive data is handled securely and in compliance with data protection regulations. Implement appropriate security measures and anonymize or pseudonymize personal information as needed.

#### 10. Data Quality Assurance:

- Implement data quality checks to maintain data quality over time. Regularly monitor the data for changes or anomalies that may affect its integrity.

#### 11. Version Control:

- Implement version control to track changes made to the dataset. This is especially important when multiple individuals are involved in data preparation.



#### 12. Data Storage:

- Store the prepared dataset in a secure and organized manner. Ensure that authorized team members can easily access and retrieve the dataset when needed.

#### 13. Data Backups:

- Regularly back up the dataset to prevent data loss due to unexpected events like system failures or accidental deletions.

#### 14. Accessibility:

- Make the dataset easily accessible to relevant team members and tools. Consider using data warehousing or database solutions to efficiently manage and access the data.

By following these steps and considering these important factors, you can effectively prepare a raw dataset that meets the needs of your organization, ensuring that the data is accurate, reliable, and ready for analysis to support informed decision-making.

housing	loan	month	deposit	duration	poutcom	y	age	mean	std.	norm.dis
no	no	july	\$ 1,200.00	261	nonexist	yes	22	43.2542	8.75654	0.00239
no	no	july	\$ 890.00	149	nonexist	yes	24			0.00406
yes	no	july	\$ 1,300.00	226	nonexist	yes	24			0.00406
no	no	july	\$ 2,800.00	151	failure	yes	25			0.00519
no	yes	july	\$ 1,700.00	307	failure	yes	25			0.00519
no	no	july	\$ 1,200.00	198	nonexist	yes	25			0.00519
no	no	july	\$ 1,600.00	139	nonexist	yes	26			0.00654
no	no	july	\$ 1,780.00	217	nonexist	yes	26			0.00654
yes	no	july	\$ 5,200.00	380	nonexist	yes	27			0.00814
yes	no	july	\$ 6,200.00	50	failure	yes	27			0.00814
no	no	july	\$ 7,600.00	55	nonexist	yes	27			0.00814
yes	no	july	\$ 1,200.00	222	nonexist	yes	27			0.00814
no	yes	july	\$ 990.00	137	success	yes	27			0.00814
yes	no	july	\$ 1,320.00	293	success	yes	28			0.00999
yes	no	july	\$ 2,800.00	146	success	yes	28			0.00999
yes	yes	july	\$ 4,750.00	174	nonexist	yes	28			0.00999
yes	no	july	\$ 1,200.00	312	nonexist	yes	28			0.00999
yes	yes	july	\$ 1,600.00	440	failure	yes	29			0.01211
yes	yes	july	\$ 1,790.00	353	failure	yes	29			0.01211
no	no	july	\$ 5,200.00	195	failure	yes	29			0.01211
no	no	july	\$ 6,200.00	38	failure	yes	29			0.01211
yes	no	july	\$ 7,600.00	262	failure	yes	30			0.01449
yes	no	july	\$ 200.00	342	failure	yes	30			0.01449
yes	no	july	\$ 890.00	181	failure	yes	30			0.01449
yes	no	july	\$ 1,300.00	172	nonexist	yes	30			0.01449
no	yes	july	\$ 800.00	99	nonexist	yes	30			0.01449
yes	no	july	\$ 1,700.00	93	nonexist	yes	30			0.01449
no	no	july	\$ 1,200.00	233	nonexist	yes	30			0.01449
no	no	july	\$ 3,600.00	255	nonexist	yes	30			0.01449
unknown	unknown	july	\$ 1,780.00	362	failure	yes	30			0.01449
no	no	july	\$ 5,200.00	348	failure	yes	31			0.01711
yes	no	july	\$ 8,200.00	386	failure	yes	31			0.01711

## 1. Probability Distribution:

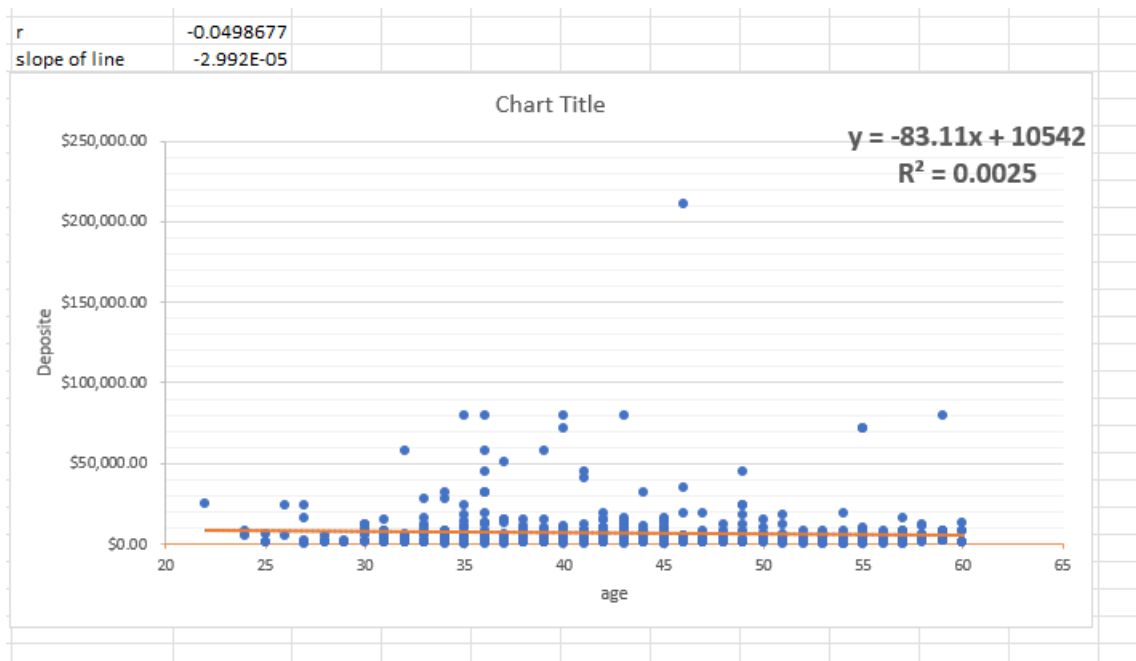
Apply probability distribution analysis to understand the distribution of account balances within different age groups. You can use techniques like the Kernel Density Estimation (KDE) to visualize the probability density function of balances for various age categories.

## 2. Regression Analysis

Perform linear or non-linear regression analysis to model the relationship between age and account balance. This can help you quantify the impact of age on account balances and make predictions based on age.

### 3. Correlation Analysis

Calculate correlation coefficients (e.g., Pearson, Spearman) to quantify the strength and direction of the relationship between age and balance. A positive correlation indicates that as age increases, account balances tend to increase as well.



**Analysis of the relationship between age and account balance is crucial for these reasons**

1. Understanding Customer Behavior: By analyzing the correlation, probability distribution, and regression relationship between age and account balance, financial institutions gain insights into customer behavior. This understanding can inform product development and marketing strategies.

2. Personalization: It allows organizations to tailor their services and products to different customer segments. For example, they can design age-specific financial products or create marketing campaigns targeting specific age groups based on their balance and savings behaviors.

3. Risk Assessment: Financial institutions can use these analyses to assess the risk associated with different age groups. For example, they can determine the likelihood of high-value deposits or the probability of customers defaulting on loans based on their age and balance.

4. Strategic Decision-Making: Insights from these analyses help in making informed decisions about business strategies, such as expanding services, managing risk, and allocating resources.

5. Customer Retention: Understanding the age-balance relationship allows organizations to develop strategies to retain and engage customers. For example, they can offer loyalty programs or financial advice tailored to a customer's life stage.

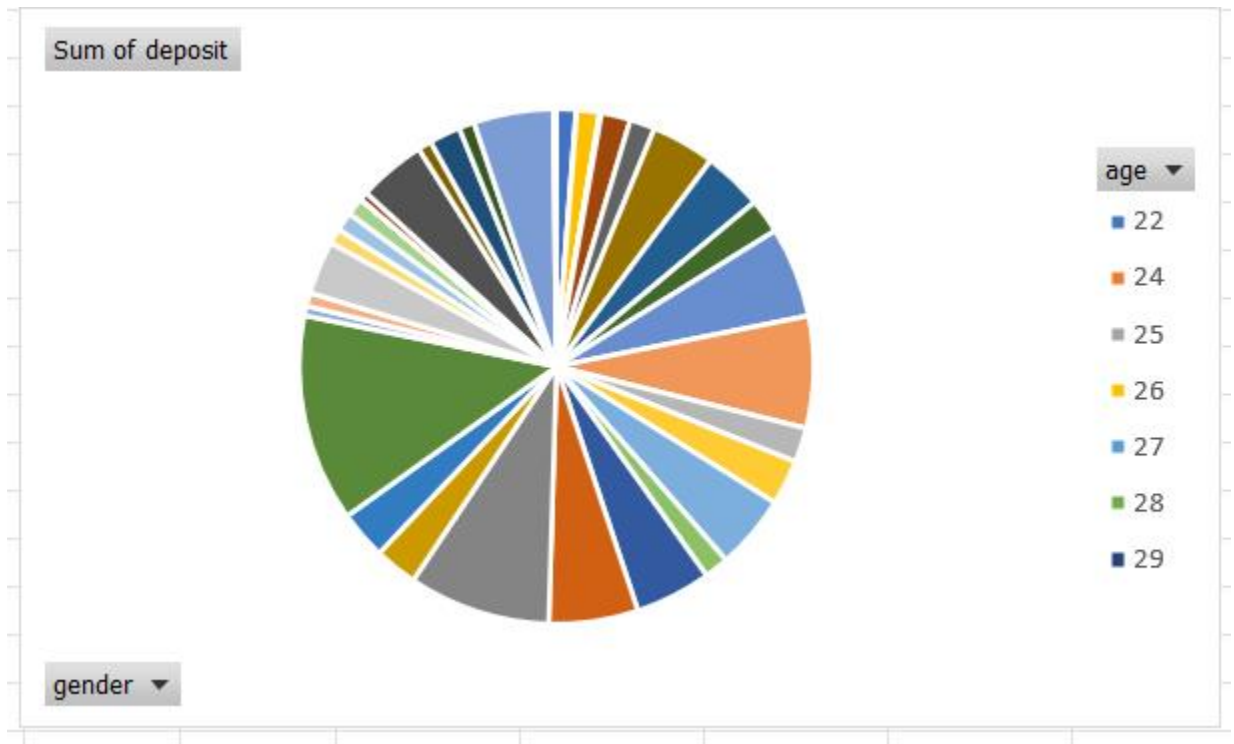
6. Data-Driven Marketing: Marketers can use this analysis to target their advertising efforts more effectively. They can craft messages that resonate with specific age groups based on their financial needs and aspirations.

7. Financial Planning: Customers can benefit from these analyses by receiving more personalized financial advice and product recommendations that align with their age and financial goals.

In essence, the analysis of the relationship between age and account balance helps organizations make data-driven decisions, enhance customer satisfaction, manage risks, and tailor their financial products and services to meet the diverse needs of their customer base.

#### Recommendations different audiences

**Customer Segmentation:** Utilize the age-balance regression insights to create more refined customer segments based on age groups. This allows the organization to offer tailored financial services and products to different customer categories. **Marketing Strategies:** Craft marketing campaigns that leverage the age-balance relationship. For example, emphasize long-term financial planning and investment opportunities for older customers, and promote savings and financial education for younger customers. **Product Development:** Consider developing new financial products that cater specifically to different age groups, capitalizing on their distinct financial needs and behaviors. **Risk Assessment:** Use the regression insights to refine risk assessment models. Recognize that age can be a factor in predicting and mitigating financial risks and defaults. **Data Collection and Privacy:** Ensure that data collection practices align with privacy regulations, particularly when collecting and analyzing customer age and balance data. Implement robust data security measures to safeguard sensitive information.



### Why we need visualized data

Visualized data is important in bank fields and industries for these reasons

1. Enhanced Comprehension: Visual representations of data, such as charts, graphs, and infographics, make complex information more accessible and easier to understand. People can quickly grasp trends, patterns, and insights that might be challenging to discern from raw data.

2. Effective Communication: Visualizations are a powerful communication tool. They enable data analysts to convey their findings to non-technical audiences, such as stakeholders and decision-makers, in a clear and engaging manner.

3. Identifying Patterns and Trends: Visualizations can reveal hidden patterns, trends, and correlations in data that might not be apparent when examining raw numbers. This is particularly valuable for data-driven decision-making.

4. Data Exploration: Visualizations facilitate data exploration by allowing users to interact with the data. Through interactive charts and dashboards, users can filter, drill down, and manipulate data to gain deeper insights.

5. Storytelling: Data visualizations can tell a story. By arranging data in a logical sequence and using visual elements to support the narrative, they can convey a compelling message or argument.

6. Decision Support: In business and decision-making contexts, visualized data helps stakeholders make informed choices. Whether it's evaluating marketing performance, financial data, or operational metrics, visualizations enable quicker and more confident decisions.

7. Early Detection of Anomalies: Data visualizations make it easier to spot outliers and anomalies. This is particularly important in fields like healthcare and finance, where early anomaly detection can prevent issues or identify opportunities.

8. Monitoring and Reporting: Visualizations are useful for real-time monitoring and reporting. They provide a snapshot of current conditions, helping organizations stay agile and responsive.

9. Comparative Analysis: Visualizations facilitate comparisons between different data points or time periods. They allow users to assess changes, differences, and similarities at a glance.

10. Aid in Problem-Solving: When faced with complex problems, data visualizations can assist in breaking down the problem into smaller, more manageable components. This can lead to more effective problem-solving and solution development.

11. Accessibility: In an era of big data, data visualization tools enable individuals to make sense of large datasets. They allow users to see the bigger picture and zoom in on specific details as needed.

12. Improved Memorization: Visual content is often more memorable than raw data. People tend to retain information better when it's presented in a visual format.



13. Engagement and Collaboration: Visualizations can engage and involve team members, fostering collaboration and discussions around data. They provide a shared point of reference for discussions and decision-making.

14. Data Quality Assessment: Visualizations can reveal data quality issues such as missing values, duplicates, or inconsistencies, making it easier to address data cleaning and normalization.

In summary, visualized data is essential for conveying information effectively, making data-driven decisions, and gaining deeper insights from complex datasets. It's a crucial tool in various domains, including business, science, healthcare, and public policy, among others.

## Evaluation

To present concise, valid, and relevant conclusions and recommendations from an accurate evaluation of the effectively cleansed and formatted banking dataset analysis that illustrates the regression relationship between age and balance, consider the following steps tailored for different audiences and organizational purposes:

### 1. Descriptive Summary

- For Executives and Decision-Makers:
- Start with an executive summary that highlights key insights.

- Conclude that there is a statistically significant relationship between age and balance in the dataset.

- Summarize the direction and strength of the relationship (positive/negative and strong/weak).

- For Data Analysts and Researchers:

- Provide detailed statistical measures of central tendency and dispersion for age and balance.

- Include correlation coefficients or regression analysis statistics to demonstrate the relationship's strength.

## 2. Data Visualization

- For Executives and Decision-Makers:

- Create visually appealing graphs, like scatter plots, that illustrate the relationship between age and balance.

- Emphasize the key trends and any outliers.

- For Data Analysts and Researchers:

- Provide detailed regression plots, residual plots, and other relevant visualizations to support the analysis.

## 3. Conclusions

- For Executives and Decision-Makers:

- Conclude that understanding the age and balance relationship can inform marketing and product development strategies.

- Recommend considering age-related factors when designing targeted banking products and campaigns.

- For Data Analysts and Researchers:

- Summarize your findings from the regression analysis, including coefficients, p-values, and R-squared values.

- Discuss the significance and practical implications of the age and balance relationship.

#### 4. Recommendations

- For Executives and Decision-Makers:

- Recommend creating tailored marketing campaigns or product offerings for different age groups based on the observed trends.

- Suggest further market research to explore how customer age impacts banking behaviors.

- For Data Analysts and Researchers:

- Suggest additional analyses or research questions related to age and balance.

- Encourage exploring additional variables that might influence this relationship.

## 5. Data Integrity and Reliability

- Emphasize that the dataset has been thoroughly cleansed and formatted, and the results are robust and reliable.

## 6. Data Accessibility

- Provide information on how to access and use the dataset and analysis for the organization's various departments and stakeholders.

By tailoring your conclusions and recommendations to different audiences and organizational purposes, you can ensure that the insights are relevant, actionable, and effectively meet the organization's needs.

Big data analysis is of paramount importance in today's data-driven landscape. It enables organizations to make informed decisions, gain a competitive edge by identifying trends and customer preferences, improve operational efficiency, manage risks effectively, drive innovation, and enhance research and development. Additionally, it supports predictive analytics, personalization, supply chain optimization, and public policy decisions, making it vital across various sectors, from healthcare and finance to environmental initiatives and security. Big data analysis empowers organizations to leverage data for increased productivity, customer engagement, and societal well-being while addressing the challenges of a rapidly evolving world.

## REFERENCE LIST

Prasad, E. (2021). *Statistics: Measures of Central Tendency*. [online] featurepreneur. Available at: <https://medium.com/featurepreneur/statistics-measures-of-central-tendency-32ffd17c5293> [Accessed 21 Sep. 2023].

Karaminis, T., Cicchini, G.M., Neil, L., Cappagli, G., Aagten-Murphy, D., Burr, D. and Pellicano, E. (2016). Central tendency effects in time interval reproduction in autism. *Scientific Reports*, 6(1). doi:<https://doi.org/10.1038/srep28570>.

Bhandari, P. (2020). *Central Tendency | Understanding the Mean, Median and Mode*. [online] Scribbr. Available at: <https://www.scribbr.com/statistics/central-tendency/>.  
admin (2019). *Modern Bank Marketing - A Comprehensive Guide (2021)*. [online] Mediaboom. Available at: <https://mediaboom.com/news/bank-marketing/>.

Emerald.com. (2019). *International Journal of Bank Marketing | Emerald Insight*. [online] Available at: <https://www.emerald.com/insight/publication/issn/0265-2323>.

Khorana, A., Pareek, A., Ollivier, M., Madjarova, S.J., Kunze, K.N., Nwachukwu, B.U., Karlsson, J., Marigi, E.M. and Williams, R.J. (2022). Choosing the appropriate measure of central tendency: mean, median, or mode? *Knee Surgery, Sports Traumatology, Arthroscopy*, 31. doi:<https://doi.org/10.1007/s00167-022-07204-y>.

StudySmarter UK. (n.d.). *Measures of Central Tendency: Definition & Examples*. [online] Available at: <https://www.studysmarter.co.uk/explanations/psychology/data-handling-and-analysis/measures-of-central-tendency/>.

