Statistical Modelling and Analysis Group Coursework

TEAM G

Om Sadigale (w1943544)

Abdullahi Mohamed Jama (w1948671)

Victor Adedayo (w1853430)

University of Westminster

BSc Data Science and Analytics

4DATA001W Statistical Modelling and Analysis

Contents

[Part 1: Data Ethics 3](#_Toc131735006)

[1.1. Victor Adedayo 3](#_Toc131735007)

[1.2. ABDULLAHI   MOHAMED JAMA 7](#_Toc131735008)

[1.3. Om Sadigale 9](#_Toc131735009)

[Part 2: BASEBALL PLAYERS DATA ANALYSIS 12](#_Toc131735010)

[Task 3: IRIS FLOWERS DATA ANALYSIS 28](#_Toc131735011)

[References: 35](#_Toc131735012)

# Part 1: Data Ethics

## Victor Adedayo

1. Introduction

Data ethics are critical principles that guide the responsible and ethical use of data across various fields. With the advent of big data and artificial intelligence, the collection, analysis, and use of data are becoming increasingly prevalent and sophisticated. The importance of data ethics lies in their ability to ensure that data is collected, analysed, and utilized in a manner that respects individual rights, privacy, and dignity while promoting social good and justice. This essay focuses on issues related to data collection, Dissemination and interpretation, including the impact of individual profiling on people's lives, the consequences of data breaches. Additionally, the role of the Human Rights Act and UK/EU legal framework in upholding data ethics is explored. Finally, this essay will provide a conclusion on the importance of data ethics and its potential impact on society.

1. The main body
   1. Issue relating to data collection.

A recent study found that over 50% of customers have no idea how their data and resources are being used, sparking debates about the ethics of data collection. Data ethics refer to the moral obligations of collecting, protecting, and utilizing personally identifiable information and its impact on individuals. Even if one is not responsible for implementing tracking code, managing a database, or training machine learning algorithms, being aware of data ethics can help identify instances of unintentional, unethical data collection, storage, or usage within an organization.

Prior to collecting personal data, businesses should consider the following data privacy ethics and their consequences. (Importance of Ethical Data Collection - prompt cloud, 28/04/2022)

* 1. Individual profiling

profiling, which is an integral part of criminal investigations, is referred to by various names such as criminal investigation analysis, crime scene analysis, psychological profiling, behavioural evidence analysis, biopsychosocial profiling, psychosocial profiling, investigative process management, criminal profiling, psychological criminal profiling, criminal behavioural profiling, offender profiling, and criminal personality profiling. Profiling can provide a deeper understanding of crime scene investigations, as the behaviour of an offender is indicative of their underlying psychological processes. Additionally, the appearance of a crime scene can reveal significant information about the perpetrator's psychopathy, sociopathy, psychopathology, or enduring character traits. Profiling is also a valuable tool for identifying subtle similarities in serial crimes. ("Profiling, Ethical Issues ." World of Forensic Science. . Encyclopedia.com., 2023 )

* 1. Impacts of data breaches on Data ethics

A data breach can result in a significant ethical consequence - the loss of privacy. Many companies possess confidential information that, if exposed, could compromise the privacy of individuals. However, a data breach not only puts the privacy of individuals at risk but also jeopardizes the entire organization and all its records, communications, and other sensitive information. (What are the Consequences of a Data Breach? EasyDMARC’s , 2022)

Data breaches can have severe implications on data ethics, which are the principles and guidelines that govern the responsible handling and use of data. One of the most significant impacts of data breaches on data ethics is the loss of privacy. When sensitive information is exposed, it can put individuals at risk, as well as potentially harm the reputation of the organization responsible for protecting the data.

Data breaches can also lead to legal and financial consequences. Organizations may face fines and lawsuits as a result of data breaches, which can be costly and damaging to their reputation. In some cases, data breaches may even result in the closure of the organization responsible for the breach.

To mitigate the impact of data breaches on data ethics, organizations must take proactive measures to protect their data. This includes implementing robust security measures, regularly monitoring for threats, and providing ongoing training and education for employees on best practices for data privacy and security. By taking these steps, organizations can protect their data and uphold the principles of data ethics. (Chen, 2018)

1. 1. The role of the human right act and uk/eu legal framework on data ethics

The Human Rights Act (HRA) and the UK/EU legal framework play a critical role in ensuring that data ethics are upheld in the handling and use of data. The HRA, which came into effect in the UK in 1998, sets out the fundamental rights and freedoms that must be respected by public authorities and bodies. This includes the right to privacy, which is particularly relevant to data ethics, as personal information is often sensitive and must be protected from unauthorized access and use. (Amelung Nina, 31 October 2020)

* 1. Dissemination and interpretation of results

The dissemination and interpretation of research results are not just technical aspects of conducting research. They also carry ethical implications, especially when it comes to data ethics. Data ethics involves considering the ethical implications of how data is collected, stored, analysed, and disseminated. Proper data ethics requires researchers to ensure that they collect data in a way that respects individual privacy, and that they analyse and disseminate the data in a way that does not cause harm to individuals or groups.

Conclusion

one of the most significant findings to emerge from this study is that data breaches have significant impacts on data ethics, including the loss of privacy, erosion of trust, and legal and financial consequences. To address these impacts, organizations must take proactive measures to protect their data and uphold the principles of data ethics. The Human Rights Act and the UK/EU legal framework play a critical role in ensuring that personal information is handled ethically and with respect for individual rights and freedoms. By following these legal standards and guidance, organizations can establish robust data protection policies and mitigate the risks of data breaches. However, there are still issues related to data collection, such as the collection of data without consent or the collection of irrelevant data, which must be addressed to ensure that data ethics are upheld in all stages of data handling. Overall, upholding data ethics is essential to protecting personal information, maintaining trust, and ensuring that data is used ethically and responsibly.

Bibliography

"Profiling, Ethical Issues ." World of Forensic Science. . Encyclopedia.com. (2023 , feb 22).

ABBAS, A. S. ( 4, Jun - Jul 2017; 11 - 20). SIDE EFFECTS OF COMPUTER ANTI-VIRUSES.

Amelung Nina, G. ( 31 October 2020). *Modes of Bio-Bordering*.

Carlos Luevanos, J.-h. Y. (: 29 March 2018). *Analysis on the Security and Use of Password Managers*.

Chen, E. T. (2018). *Ethical Issues and Considerations of Big Data*. Chapman and Hall/CRC. ENGST, G. F. (2009). *take control of your wi fi security* . tidBITS.

Floriano De Rango, D. C. (Received 10 October 2005; Revised 2 April 2006; Accepted 13 June 2006). *Wireless Communications and Networking*. Hindawi.

Importance of Ethical Data Collection - prompt cloud. (28/04/2022).

Mamaghani, F. (1 March 2002 ). Information Management & Computer Security.

Moran, J. O. (2010). *Protecting Your Data. In: Getting StartED with Windows 7. Apress. https://doi.org/10.1007/978-1.*

nelson, s. (2011). *pro data backup and recovery .*

Schaub, C. S. (2020). Interactive Stories for Security Education: A Case Study on Password Managers.

What are the Consequences of a Data Breach? EasyDMARC’s . (2022, 06 16).

## ABDULLAHI   MOHAMED JAMA

Data collection has various ethical concerns, especially when dealing with sensitive information. The Cambridge Analytica scandal in 2018 is a good example, where the company collected data from many Facebook users without their consent for political purposes. This highlights the need to get permission from individuals before gathering their data and ensuring it's used only for the intended purpose. Big data can lead to individual profiling, which might result in discrimination or exclusion. Predictive policing algorithms are an example, using historical data to predict future criminal activities. These algorithms can cause biases against specific racial or ethnic groups, reinforcing discrimination. To address this, we need to make sure that algorithms are transparent and regularly checked for biases. Data breaches are another concern, as they can have severe consequences for both people and organisations. The 2017 Equifax breach, where hackers accessed personal information of millions of consumers, shows the importance of strong security measures to protect sensitive data. Organisations must have solid data protection policies, and governments need to enforce strict regulations to hold them accountable. Interpreting big data analytics can have significant effects, especially when it comes to public policies or decision-making processes. Using algorithms to determine welfare benefits, for example, might result in unfair distribution or denial of services to vulnerable populations. To reduce these risks, it's important to involve various stakeholders, including data scientists, policymakers, and affected communities, in interpreting and sharing results. The Human Rights Act (1998) and the UK/EU legal framework play a critical role in safeguarding individuals' rights when it comes to big data. The General Data Protection Regulation (GDPR) from 2018 is a key example, giving people more control over their personal data and imposing penalties on organisations that don't comply. To ensure ethical use of big data, we need to keep strengthening and adapting these legal frameworks to deal with new challenges and technology advancements. To sum up, the ethical issues surrounding data collection and big data analytics involve concerns about informed consent, individual profiling, data breaches, and the dissemination and interpretation of results. As we rely more on data-driven decision-making, it's essential to tackle these ethical challenges to avoid harming people and communities. The Human Rights Act and the UK/EU legal framework, including the GDPR, offer important safeguards for individuals' rights. However, continuous efforts are needed to adapt and strengthen these legal frameworks in response to emerging challenges and technological advancements. Collaboration between multiple stakeholders, such as data scientists, policymakers, and affected communities, is necessary to ensure the ethical use of big data and create a more equitable and just society.

References:

Cadwalladr, C., & Graham-Harrison, E. (2018). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. The Guardian. Retrieved from<https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election>

Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.

Ferguson, A. G. (2017). The rise of big data policing: Surveillance, race, and the future of law enforcement. NYU Press.

Krebs, B. (2017). The Equifax breach: What you should know. Krebs on Security. Retrieved from<https://krebsonsecurity.com/2017/09/the-equifax-breach-what-you-should-know/>

O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown.

Voigt, P., & von dem Bussche, A. (2017). The EU General Data Protection Regulation (GDPR): A practical guide. Springer International Publishing.

## Om Sadigale

With increased digitalization across all industrial sectors, businesses are gathering and using more data to better understand their clients' demands and make client-centred decisions.

181 zettabytes of data are anticipated to be produced by 2025. Due to the data's exponential growth, most businesses are now focusing on the operational aspects of data management, including data collection, storage, and the involvement of data scientists in the management of data. However, it is crucial to understand how data can be acquired, kept, and used ethically without endangering the privacy of persons in addition to the operational aspects of data management. Data ethics is the assessment of data-related procedures, such as the gathering, storing, using, and publishing of data in an ethical manner so that people's confidence is maintained.

The topic of data ethics is whether data handling is unethical. The client has a vote in how the data will be used and shared throughout, whether the data is acquired internally or outside from third-party suppliers or partners. Data ethics aids in decision-making around data usage, preventing data misuse that might cost the organisation money and damage its brand.

The way the data is gathered may be of issue. The ethical problems with data collecting that we run into are as follows:

Lack of Consent: Before collecting any information, the persons' consent must be obtained. To deliver customer-focused advertising, the authorization may be obtained by accepting the cookies or by signing a consent form.

Uncertain about how data will be used: The participant should be made aware of the goal of the data collection to gain their consent

Data security breach: Participants' names should remain hidden if they selected anonymity when giving their assent. Other problems include abusing data to take advantage of people, gathering information unrelated to the subject of the survey, or collecting data in an objective manner. (*StudySmarter UK*)

Individual profiling is the processing of personal data to look at a candidate's personality and make predictions based on that analysis. Regarding data ethics, individual profiling presents a challenge because there needs to be a legal foundation for it. Profiling can be used to foretell a person's financial situation or physical condition. (*Dataetisk Tænkehandletank* (2021))

To prevent data breaches and unauthorised access to private data, the obtained data must be preserved securely. The primary effects of data breaches include financial losses, lost income, a decline in share price as a result of investors' and consumers' diminished faith, costs incurred to recompense customers, fines and penalties, investments in new security systems, etc. Yahoo had suffered data breach and had to sold at discounted rate to Verizon.

The Reputation loss due to data breaches is damaging to company as the customers and investors lose trust in company and stop doing business. Also, if becomes difficult to get the new customers and investors. Business disruption occurs after data breach till the complete investigation is made and this will cause the downtime for operational resulting in financial losses.

In addition to the financial costs to the business, the legal ramifications of not protecting client information could result in identity theft, fines from the government, and compensation claims from customers. Equifax's 2017 data leak led to the payment of $700 million in compensation to US consumers. (MacKay, J. (2023))

Once the data is collected ethical, it is very important to disseminate the data to users such as government officials, researchers or general public based on the national rules and regulations. The data made available to users should not compromise the confidentiality of participants or organization. The results from profiling or survey should not biased. The interpretation of data should be done considering the short term and long-term effect of it and should be presented in respectful way considering the culture, or nature of data.

The Human Rights Act states the right to respect the private and family life thus the public bodies should respect the personal life of individual and any personal information related to them. The personal information stored should be justifiable and should not be misused. The UK Data Ethic framework guides the government and the public sector on the appropriate and responsible data use. The three overarching principles such as transparency, accountability and fairness are applicable throughout the entire project. (*Data Ethics Framework*)

Reference:

Contact us Preeti Shivpuri Omnia AI pshivpuri@deloitte.ca 416-874-3385 (2022) *The rising importance of Data Ethics: Deloitte Canada*, *Deloitte*. Available at: https://www2.deloitte.com/ca/en/pages/deloitte-analytics/articles/the-rising-importance-of-data-ethics.html (Accessed: April 5, 2023).

Edquist, A. *et al.* (2022) *Data ethics: What it means and what it takes*, *McKinsey & Company*. McKinsey & Company. Available at: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/data-ethics-what-it-means-and-what-it-takes (Accessed: April 5, 2023).

Cote, C. (2021) *5 principles of data ethics for business*, *Business Insights Blog*. Available at: https://online.hbs.edu/blog/post/data-ethics (Accessed: April 5, 2023).

*Data Ethics Framework* (no date) *Government Digital Service*. Available at: This publication is also available on our website at www.gov.uk/government/ publications/data-ethics-framework (Accessed: April 5, 2023).

*Ethical issues in data collection* (no date) *StudySmarter UK*. Available at: https://www.studysmarter.co.uk/explanations/english/research-and-composition/ethical-issues-in-data-collection/ (Accessed: April 5, 2023).

*FAQ - principles of data ethics · Dataetisk Tænkehandletank* (2021) *Dataetisk Tænkehandletank*. Available at: https://dataethics.eu/faq-principles-of-data-ethics/#:~:text=Profiling%20is%20the%20automatic%20processing,work%2C%20financial%20situation%20or%20health. (Accessed: April 5, 2023).

MacKay, J. (2023) *5 damaging consequences of a data breach*, *MetaCompliance*. Available at: https://www.metacompliance.com/blog/data-breaches/5-damaging-consequences-of-a-data-breach (Accessed: April 5, 2023).

# Part 2: BASEBALL PLAYERS DATA ANALYSIS

Contributors:

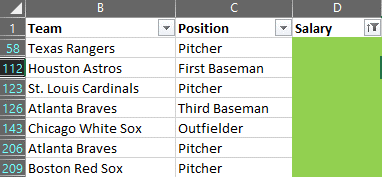
* Om worked on the solving the question.
* Abdullahi and Victor worked together to check for errors

The Data Analysis was conducted in Excel.

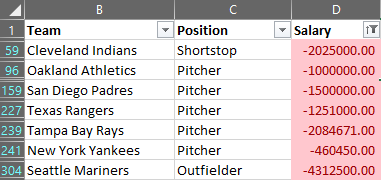
Further results were conducted from the database provided ‘Baseball\_Salaries’.

1. The erroneous entries in the Salary data:

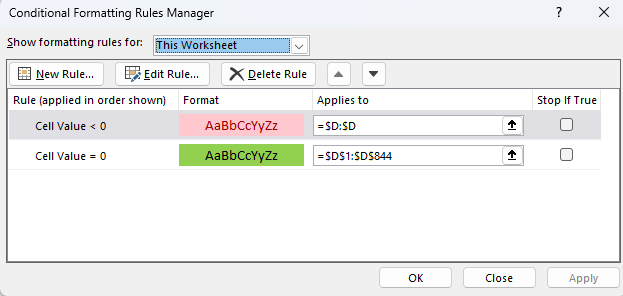
* Empty Data Sets



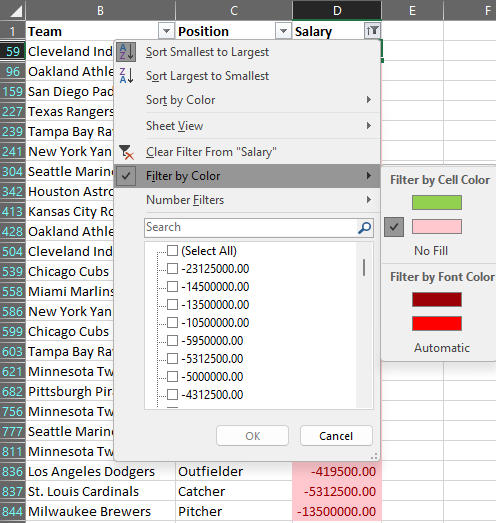
* Negative Data Values



This was done by using the function Conditional formatting to create new rule

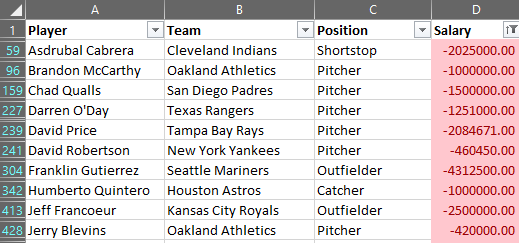


* Next is to select the Data option in the menu bar and selecting Filer under the sort & Filter. This will allow the user to separate the values.



* To Identify the errors, they were marked by colors to differentiate.





* No fill option gives a clean data set with the errors removed

1. Frequency Histogram of Salary for each position

There are 9-difference position mentioned in the data sets. The Frequency function was used to calculate frequency for the graph. (www.youtube.com. (n.d.).)

Example:

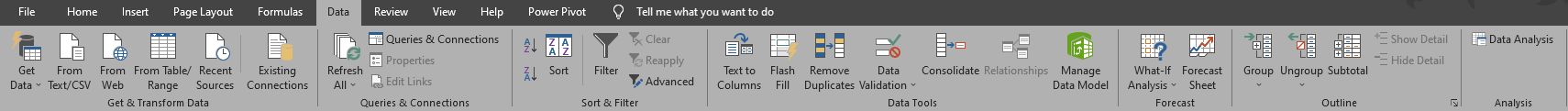


The Bins were calculated by resolving the difference between the max and min data values and divided it by the number if bins requested. Upon selecting Bins the ranges were decided.

Example:

There is another method to solve this question, which was by using the Data Analysis function under Analysis tool which can be found at corner of the Data section.



This function can create an histogram, frequency and range values by using the data such as bins and salary data values.

Frequency Histogram for Catcher:

Frequency Histogram for Designated Hitter:

Frequency Histogram for First Baseman:

Frequency Histogram for Infielder:

Frequency Histogram for Outfielder:

Frequency Histogram for Pitcher:

Frequency Histogram for Pitcher:

Frequency Histogram for Shortstop:

Frequency Histogram for Third Baseman:

There were no systematic errors found which was conducted using the trendline (line of best fit). If the line passes through 0, it implies that there is a systematic error.

Based on the graph it can be stated that all positions except for Designated Hitter have a Negative line of best fit for salary. From the negative trendline it can be implied higher the salary lower the frequency (players). For Positive trendline has the relation of lower the salary lower the frequency.

Negative Trendline:

Positive Trendline:

1. Standard derivation per position and as a whole

The most convenient and fastest way to complete this question was to use the function Standard Deviation (Sample). (EDUCBA. (2019))

Example:



An alternate method is to use the traditional technique (manually) which is calculating the mean, variance, standard deviation, deviation from the mean and Square Deviation. This method was used for the first two position to prove that it works.

Calculating for Catcher

* The first step is to calculate the sum.



* The second step is to find the average/mean. Divide the sum by the number of data values.



* Third step is to find the Deviation from the mean. Subtracting each value with the mean.



* Calculate the total sum of the values from Deviation from the mean which should add up to zero because the number of negatives and positivises will be same.



* Next step is to find the Square Deviation. Square the values from Deviation from the mean.



* Calculate the total sum of the Square Deviation values.



* Using the total sum of Square Deviation calculate the variance by dividing the total sum by the number of data values



* Using the data from variance calculate the square root of it which will give the solution for Standard Deviation



First Technique Results

|  |  |  |
| --- | --- | --- |
|  | Position | Standard Deviation (Sample) |
|  | Catcher | 3610429.01 |
|  | Designated Hitter | 4758319.488 |
|  | First Baseman | 6341208.893 |
|  | Infielder | 6902662.901 |
|  | Outfielder | 5192324.589 |
|  | Pitcher | 4077895.375 |
|  | Second Baseman | 3321025.53 |
|  | Shortstop | 3601789.809 |
|  | Third Baseman | 6002060.053 |
| Whole Standard Deviation (Sample) | | 4522792.214 |

Alternative technique results

|  |  |  |
| --- | --- | --- |
|  | Position | Standard Deviation (Sample) |
|  | Catcher | 3610429.01 |
|  | Designated Hitter | 4758319.488 |

1. Illustrate the number and proportion of players per position:

|  |  |  |
| --- | --- | --- |
|  | Percentage of Total | Absolute Values |
| Catcher | 8.387096774 | 65 |
| Designated Hitter | 0.985915493 | 7 |
| First Baseman | 5.263157895 | 37 |
| Infielder | 0.750750751 | 5 |
| Outfielder | 21.48260212 | 142 |
| Pitcher | 72.44701349 | 376 |
| Second Baseman | 39.16083916 | 56 |
| Shortstop | 52.87356322 | 46 |
| Third Baseman | 100 | 41 |

Please refer attached:



# Task 3: IRIS FLOWERS DATA ANALYSIS

Contributor:

* Abdullahi and Victor worked together to answer the questions
* Om worked on the errors

Introduction

The description of the Edgar Anderson's Iris dataset. "This famous (Fisher's or Anderson's) iris data set gives the measurements in centimetres of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of three species of iris. The species are Iris setosa, versicolor, and virginica."

1. Present Iris Data

It was necessary to identify the variable names, their statistical classes, and the size of the vectors. The data values displayed below are the first 6 lines of values for each variable and the method used for that was printing the first 6 rows of the dataset in R using head(iris) command. (Schork, J. (2021))

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Sepal Length | Sepal Width | Petal Length | Petal Width | Species |
| 1. | 5.1 | 3.5 | 3.5 | 1.4 | Setosa |
| 2. | 4.9 | 3.0 | 3.0 | 1.4 | Setosa |
| 3. | 4.7 | 3.2 | 3.2 | 1.3 | Setosa |
| 4. | 4.6 | 3.1 | 3.1 | 1.5 | Setosa |
| 5. | 5.0 | 3.6 | 3.6 | 1.4 | Setosa |
| 6. | 5.4 | 3.9 | 3.9 | 1.7 | Setosa |

summary(iris[,1:4])

Sepal.Length Sepal.Width Petal.Length Petal.Width

Min. :4.300 Min. :2.000 Min. :1.000 Min. :0.100

1st Qu.:5.100 1st Qu.:2.800 1st Qu.:1.600 1st Qu.:0.300

Median :5.800 Median :3.000 Median :4.350 Median :1.300

Mean :5.843 Mean :3.057 Mean :3.758 Mean :1.199

3rd Qu.:6.400 3rd Qu.:3.300 3rd Qu.:5.100 3rd Qu.:1.800

Max. :7.900 Max. :4.400 Max. :6.900 Max. :2.500

To answer this question, we calculated the standard descriptive statistics, including the central tendency (mean) and standard deviation for all 4. We used R functions mean () and SD () to find the mean and standard deviation for each variable, and presented the results. (Schork, J. (2021))

1. Scatter Plot Questions
2. Possible Scatter Plots

While looking for the possible scatter plots, it is necessary to find the variable n (numerical variables) and k (unique pairs).

= 4

= 2

Applying the combinations formula:

=

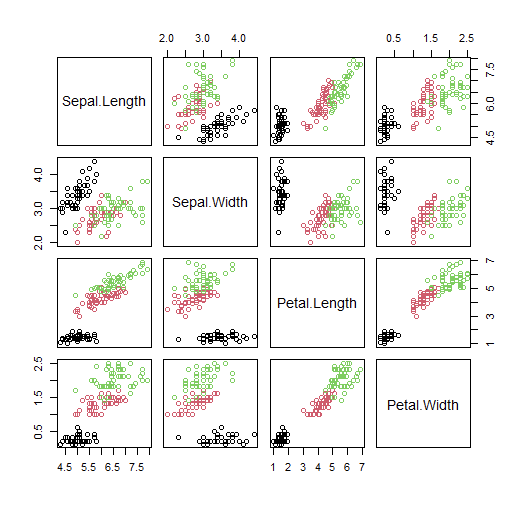
= = = 6

Conclusion:

There are 6 possible scatter plots of 4 numbers. ((n.d.). *Combination*)

1. Scatter Plots

pairs (iris [,1:4], col = iris$Species) gives the ‘i’ value. R functions like plot () or pairs() were used to create scatter plots for each unique variable pairs. (R CODER. (2020))



1. Examined the scatter plots visually to identify which plots exhibited a linear relationship. A linear relationship between two variables indicates that one variable change proportionally with the other, suggesting a potential correlation or association between them.
2. Frequency histograms for each variable:

 R's hist() function was used to create (relative) frequency histograms for each numerical variable. Set the argument probability = TRUE to display relative frequencies instead of counts. (R CHARTS)

Code:

par(mfrow=c(2,2))

Sepal Length:

hist(iris$Sepal.Length, col="lightblue", main="Sepal Length", xlab="Sepal Length (cm)", ylab="Frequency", freq = FALSE)

Sepal Width:

hist(iris$Sepal.Width, col="lightblue", main="Sepal Width", xlab="Sepal Width (cm)",

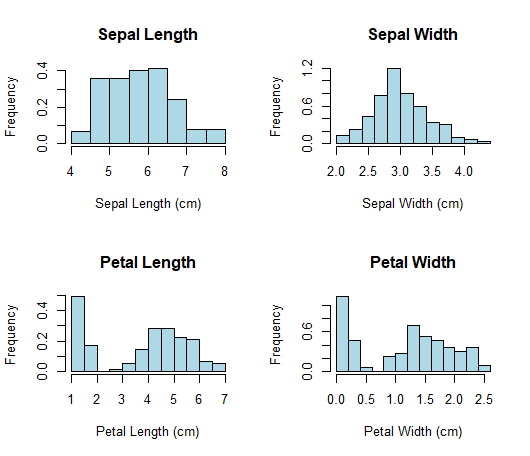
ylab="Frequency", freq = FALSE)

Petal Length:

hist(iris$Petal.Length, col="lightblue", main="Petal Length", xlab="Petal Length (cm)", ylab="Frequency", freq = FALSE)

Petal Width:

hist(iris$Petal.Width, col="lightblue", main="Petal Width", xlab="Petal Width (cm)", ylab="Frequency", freq = FALSE)



1. Normally Distributed Variable:

By visually examining the histograms, we identified the variable that appeared to be normally distributed, which we then referred to as variable X. (R CHARTS)

Sepal width appears to be normally distributed

X = Speal Width

1. The formula for model normal probability density function:

We isolated variable X and calculated its mean and standard deviation using R's mean () and sd() functions. We then defined the normal probability density function (PDF) using the formula:

=

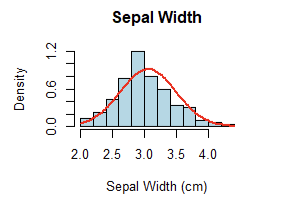
To plot the PDF, we used R's curve () function, overlaying it on the histogram to visualize the fit. (R CHARTS)

Code:

hist(X, col="lightblue", main="Sepal Width", xlab="Sepal Width (cm)", ylab="Density", freq = FALSE)

curve (dnorm (x, mean=mean\_X, sd=sd\_X), add=TRUE, col="red", lwd=2)

Histogram:



1. Using data in X to calculate the probability:

Procedure:

* First step is to assume variable X is normally distributed and calculate its mean and standard deviation.
* Next is to use R pnorm () function to estimate the probabilities for given conditions and expressed them as percentages.

The probabilities tell us the likelihood(probability) of iris flowers having Sepal.Width values within specified ranges.

1. X > 3

P\_X\_greater\_than\_3 = 1 - pnorm(3, mean = mean\_X, sd = sd\_X)

1. 2.5 < X < 3.5

P\_X\_between\_2\_5\_and\_3\_5 = pnorm(3.5, mean = mean\_X, sd = sd\_X) - pnorm(2.5, mean = mean\_X, sd = sd\_X)

1. X < 2.5

P\_X\_less\_than\_2\_5 = pnorm(2.5, mean = mean\_X, sd = sd\_X)

Z\_3 = (3 - mean\_X) / sd\_X

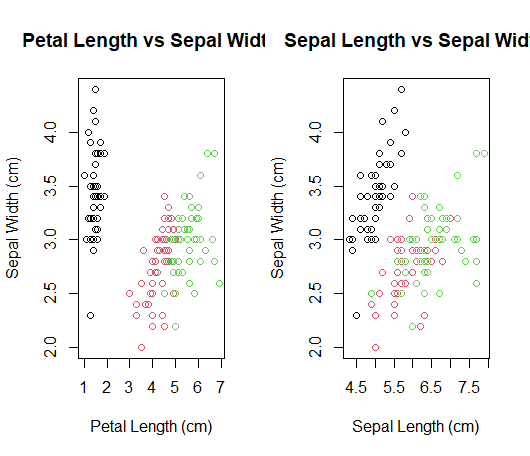
Z\_2\_5 = (2.5 - mean\_X) / sd\_X

Z\_3\_5 = (3.5 - mean\_X) / sd\_X



Procedure:

* First create scatter plots of variable X (Sepal.Width) against petal length and sepal length. It is noticeable that the data points formed two distinct clusters in each plot.
* Next is to set a threshold to separate the first cluster with lower values of petal length and width.
* The last step is to check if the data points in this cluster belonged to the same species by examining the 'Species' variable.



Code: Displays the first cluster

first\_cluster <- iris[(iris$Petal.Length <= 2.5) & (iris$Sepal.Width <= 3.3), ]



Calculated the 95% confidence interval for the true mean of variable X using all data points, by applying relevant R functions and formulas. ( GeeksforGeeks. (2022))

R Code:

n <- length(X)

mean\_X <- mean(X)

sd\_X <- sd(X)

standard\_error <- sd\_X / sqrt(n)

lower\_limit <- mean\_X - 1.96 \* standard\_error

upper\_limit <- mean\_X + 1.96 \* standard\_error

confidence\_interval <- c(lower\_limit, upper\_limit)



Created a subset, S, of X that included only the values of the species identified in the previous question, and established the 95% confidence interval for the true mean of X based on this subset.( GeeksforGeeks. (2022))

R Code:

S <- iris[iris$Species == "setosa", "Sepal.Width"]

S= [1:50]



Inferential statistics is a way to learn about a large group by looking at a smaller sample from that group. It helps us estimate important values, like the average, and gives us a range where we think the real value is. This is different from what we did in question 3, where we just guessed probabilities using a normal curve without considering how the data varied. By looking at the data's variability, inferential statistics helps us understand the uncertainty in our guesses, which is important when making decisions based on data.

# References:

Excel:

www.youtube.com. (n.d.). *Use Excel 2016 to make Frequency distribution and Histogram for quantitative data*. [online] Available at: https://www.youtube.com/watch?v=Giewd9yH4q0&ab\_channel=KwaiChan [Accessed 7 Apr. 2023].

‌

EDUCBA. (2019). *Standard Deviation in Excel (Formula, Examples) | How to Use?* [online] Available at: https://www.educba.com/standard-deviation-in-excel/.

R:

Schork, J. (2021). Introduction to Statistics Using the R Programming Language. *Statistics By Jim*. [online] 23 Jun. Available at: https://statisticsbyjim.com/basics/introduction-statistics-r-programming-language/.

Corporate Finance Institute. (n.d.). *Combination*. [online] Available at: https://corporatefinanceinstitute.com/resources/data-science/combination/.

R CODER. (2020). *SCATTER PLOT in R programming 🟢 [WITH EXAMPLES]*. [online] Available at: https://r-coder.com/scatter-plot-r/.

‌ R CHARTS | A collection of charts and graphs made with the R programming language. (2020). *Frequency histogram in R*. [online] Available at: https://r-charts.com/distribution/frequency-histogram/ [Accessed 4 Apr. 2023].

‌ R CODER. (2020). *NORMAL DISTRIBUTION in R [dnorm, pnorm, qnorm and rnorm]*. [online] Available at: https://r-coder.com/normal-distribution-r/.

GeeksforGeeks. (2022). *How to Find Confidence Intervals in R?* [online] Available at: https://www.geeksforgeeks.org/how-to-find-confidence-intervals-in-r/.