SOLENT UNIVERSITY SOUTHAMPTON

IN PARTNERSHIP WITH QA

Analytics and Business Intelligence

(QHO539)

BSc (Hons) Computing

Personal Learning Report

Academic Year 2024-2025

Name: Andre Ramos

Tutor: Ajmal Gharib

Student ID: 10300471

Word count: 2500

Date: 31/01/2025

Contents

1. Introduction

This report reflects on my experiences during the Analytics and Business Intelligence (QHO539) module, using Gibbs' Reflective Cycle to evaluate my learning journey. Throughout this module, I explored a variety of tools, concepts, and methodologies, including data governance, statistical analysis, data visualisation, and machine learning. Each session challenged me to think critically and encouraged me to develop technical skills that feel increasingly applicable to real-world scenarios.

As the weeks went on, I approached each task with curiosity, even when they felt overwhelming. Topics like machine learning algorithms stood out as particularly challenging at first. However, breaking them into smaller, more manageable steps allowed me to make sense of the complexity. Along the way, I encountered moments of frustration—especially when debugging errors or learning new tools. Despite this, I found these challenges to be valuable learning opportunities. One example that remains clear in my mind was when I struggled with data visualisation techniques. Revisiting tutorials and experimenting with various approaches eventually made a noticeable difference, and this process boosted my confidence in handling such tasks.

Reflecting on my progress, I've identified areas where I need to improve. Managing my time more effectively when tackling complex tasks is one such area. I've also recognised how important patience and persistence are when working with unfamiliar software or methodologies. Moving forward, I plan to dedicate more focused time to practising technical skills and actively seek constructive feedback from peers and instructors. These steps will not only help address my current challenges but also enhance my ability to adapt to new situations.

What excites me most about this module is its connection to real-world applications. The focus on data-driven decision-making and problem-solving has shown me how these skills can be applied in professional settings, such as improving business operations or analysing customer behaviour. These insights have deepened my understanding of the subject and given me confidence in my ability to contribute effectively to data-focused roles in the future.

Ultimately, this report captures the progress I've made in mastering the tools and concepts presented during the module. Reflecting on my development, I can see how far I've come in bridging the gap between theory and application. What stands out most is how practical and transferable these skills are—ready to be applied in professional environments where they can create meaningful impact.

2. Weekly log

2.1 Week 1

This week, I learned about the roles of statisticians, data analysts, and business analysts, particularly how they support decision-making in organisations. Statisticians work at different levels, helping organisations use data to make informed decisions (Abraham, 2005). I also found it interesting that analytics roles are not just about increasing profits but also about reducing risks and preparing for the future (Chen et al., 2012).

Data analysts use Python for data preparation, SQL for managing large datasets, and Tableau to present insights in actionable formats (Collier & Powell, 2024). Data scientists, on the other hand, require advanced programming, machine learning skills, and tools like Hadoop and Spark, and organisations often prefer candidates with a master's degree and practical experience (Smaldone et al., 2022). Business analysts need to be proficient in tools like UML and BPMN, and they often work in Agile environments (Brandenburg, 2009). These details helped me better understand the technical and professional expectations of each role.

At first, I struggled to find reliable resources about statisticians, but once I did, everything became clearer. I noticed a lot of emphasis on technical skills, but not much on soft skills like storytelling and communication. I think this is misleading because these skills are essential to explain findings effectively (Brandenburg, 2009).

Although the salaries for these roles are high, I'm more motivated by the challenge of helping organisations make better decisions. My plan is to improve my Python skills and work on storytelling projects that combine technical and soft skills. This will help me move closer to my goal of becoming a data architect.

2.2 Week 2

This week, I learnt descriptive analysis, which strengthened my grasp of statistical measurements and their application in real-world situations. Kaur, Stoltzfus, and Yellapu (2018) presented significant insights into the significance of studying data distributions and outliers, highlighting the role of statistical tools in data interpretation for better decision-making.

Working on the dataset, I explored central tendencies, dispersion metrics, and identified outliers using box plots. This process was seamless due to my familiarity with descriptive analysis. However, technical challenges arose when Excel froze, requiring a restart. These interruptions, while frustrating, tested my patience and reaffirmed my ability to recover quickly due to a clear understanding of the tasks.

Creating a dashboard was a particularly enjoyable challenge. It enhanced my ability to present data visually, ensuring insights were not just understood but also conveyed effectively. Unfortunately I struggled with producing a scatter plot that differentiated product categories due to uniform colouring, which hindered the storytelling aspect of the visualisation. This limitation has motivated me to refine my data visualisation skills.

Moving forward, my plan is to focus on mastering advanced data visualisation techniques, ensuring clarity and storytelling in all graphical outputs.

I aim to explore software alternatives to mitigate technical disruptions and enhance productivity. This week's challenges and accomplishments have inspired me to approach future analyses with an emphasis on improving both technical proficiency and visual impact.

2.3 Week 3

This week focused on descriptive analysis and data visualisation, using techniques like measures of central tendency (mean, median, mode) and dispersion (range, variance, standard deviation). We also worked on creating charts like histograms and bar graphs to summarize data, spot patterns, and present findings in a way that supports better decision-making (Field, 2018).

Feelings

These concepts felt like a natural step forward after learning them in QHO328 (Foundation Mathematics) and applying them in QHO429 (Introduction to Databases). I feel confident with basic calculations, but working with percentiles and quartiles gave me a clearer understanding of how to divide data into meaningful segments. Pie charts and bar graphs are still my go-to visuals for showing proportions and trends, especially when analysing sales data in my retail work.

Evaluation

This week reinforced that descriptive analysis is more than just calculations—it's about simplifying data to make it actionable. For example, using standard deviation to understand variability helped me see how it highlights inconsistencies in sales data. The tasks also showed why accurate and clear visualisations are key for sharing insights with stakeholders.

Analysis

Excel worked well for basic tasks, but I am starting to notice its limits when it comes to more advanced analysis. This week helped me see how descriptive statistics can make data easier to understand for non-technical audiences, which is particularly useful in my retail role (Agresti & Finlay, 2021).

Action Plan

- Practice making more advanced charts using Excel's features (Microsoft, 2024).
- Learn how to create interactive dashboards with Tableau (Tableau, 2024).
- Apply these techniques to real-world projects to find deeper insights.

- Look for resources to improve my skills in interpreting and explaining data (Few, 2023).
- Share ideas with colleagues to learn from their experiences with visualisation tools.

Statistic	Values	Formulas
Mean	120.89	=AVERAGE(L2:L9995)
Standard Deviation	31.97	=STDEV(L2:L9995)
Median	140.25	=MEDIAN(L2:L9995)
25th Percentile	99	=QUARTILE.INC(L2:L9995, 1)
75th Percentile	140.25	=QUARTILE.INC(L2:L9995, 3)

2.4 Week 4

This week focused on data preparation, emphasising how to ensure datasets are complete, accurate, and consistent. Tableau was introduced as a visualisation tool, offering interactive and real-time features that make it a strong alternative to Power BI (Few, 2023). These skills are critical for creating high-quality datasets and presenting insights effectively to support decision-making.

Feelings

Having worked with Power BI before, I noticed many similarities between the two tools. Tableau's features, like dashboards and drill-down options, were new to me but opened creative ways to share data insights. Initially, I found Power BI easier to use, but after practicing with Tableau, it felt more intuitive, particularly for creating symbol maps to analyse geographical data. Expanding my skills with Tableau has been valuable for working toward my career goal in data architecture.

Evaluation

This week reinforced how essential good data preparation is for reliable analysis. Tableau, like Power BI, is highly effective for creating visual stories that explain trends and predict outcomes. Its interactive dashboards are especially helpful for engaging non-technical audiences and simplifying complex insights (Cyntexa, 2024).

Analysis

Using Tableau showed me how to create data-driven stories that suit different audiences. This week also emphasised the importance of clean data, as errors can lead to misleading results. Tableau's ability to handle large datasets more efficiently than Power BI could make it ideal for complex projects in professional settings (Edmond and Crabtree, 2023).

Action Plan

- Practice Tableau's advanced features, like calculated fields and data blending (Noble Desktop, 2024).
- Work on real-world datasets to build interactive dashboards and explore Tableau's capabilities further (Coursera, 2024).
- Share visualisations with colleagues to gather feedback and refine my skills using Tableau's collaboration tools (Simplilearn, 2024).

2.5 Week 5

This week was about setting up Python environments and using libraries like Pandas and NumPy for data analysis. These tools are critical for handling datasets, performing descriptive analysis, and automating workflows. They offer clear advantages over Excel, especially for scalability and repeatability (McKinney, 2017).

Feelings

Having worked with Python libraries before, I found setting up the environment and importing datasets straightforward. Using the Pima Indian Diabetes dataset, I revisited statistical measures like mean, median, and standard deviation. This felt like a useful refresher and boosted my confidence in Python's ability to handle tasks much more efficiently than Excel, particularly with large datasets. Python's ability to quickly fill missing values and calculate key statistics made me realize how much time it can save.

Evaluation

Reflecting on this week, I saw how Python makes repetitive tasks faster and more accurate. Using Python with the Pima Indian Diabetes dataset showed how it scales well for real-world scenarios. Compared to Excel, Python's simplicity for automating workflows made it stand out as a key tool for data analysis.

Analysis

This week highlighted the importance of Python in creating effective data analytics workflows. It reinforced how it can manage complex datasets and automate tasks, saving time and improving accuracy. Working on real-world datasets like the diabetes data demonstrated Python's value in scaling up workflows and why it's critical for modern data analysis.

Action Plan

Work with more real-world datasets to sharpen my skills in Pandas and NumPy.

Use visualisation libraries like Matplotlib and Seaborn to create better data visualisations.

Practice automating data cleaning workflows to improve efficiency.

Build a Python-based project to show my ability to work with large datasets effectively.

2.6 Week 6

This week focused on feature selection, data preparation, and working with algorithms like K-Nearest Neighbours (KNN). These techniques are key to building accurate predictive models, and I found that feature selection plays a big role in improving model performance and accuracy.

Feelings

At first, I struggled with feature selection because choosing and prioritising the most relevant variables wasn't easy. After reviewing the material, I felt more confident applying these ideas in Python. Compared to Excel, Python made everything faster and easier to manage, especially for large datasets. Revisiting my house price prediction project using KNN reminded me how much feature selection improved the accuracy of the results.

Evaluation

This week showed me how important feature selection is, especially for KNN. My earlier project, where I achieved strong results (RMSE of £3,454.19 and R^2 of 0.9999) on a dataset of over 23,000 records, highlighted this (GitHub: Piripack/House-price-prediction).

Using Python for workflows and handling complex datasets reinforced its value, especially for automating repetitive tasks.

Analysis

KNN relies heavily on properly selected features since it doesn't adjust feature importance automatically like some other algorithms. This made me realize how important it is to master feature selection to build effective models. Python's tools, like Scikit-learn, simplify this process and make model testing and evaluation much easier.

Action Plan

- Try advanced feature selection methods like Recursive Feature Elimination.
- Explore Scikit-learn to automate feature selection and evaluate models.
- Test ensemble methods like Random Forests to compare their results with KNN.
- Document workflows to improve consistency and build reusable processes.

2.7 Week 7

Description

This week, I focused on applying both univariate and multivariate techniques to analyse datasets using statistical measures like correlation, skewness, and standard deviation. These activities were strongly related to the purpose of using statistical methods to extract actionable insights and apply them to actual settings (Solent, 2024).

Feelings

The concepts of univariate and multivariate analysis seemed basic and drew on my past knowledge. Pie charts are still my preferred option for simple datasets because of their clarity in displaying proportions. However, scatterplots and correlation matrices were more useful for analysing relationships between many variables. Having prior experience with Jupyter Notebooks in Google Colab and PyCharm sped up the code reuse process, allowing me to focus on comprehending findings rather than on how to setup the tools.

Evaluation

These techniques provided significant insights into patterns within datasets. For example, correlation matrices offered a concise way to identify relationships, while scatterplots visually reinforced these findings. In a real-world context, these tools could be applied to warehouse analysis, such as identifying an underperforming supplier whose stock movement and sales contribution are insufficient to justify ongoing costs. While Excel performed these simple tasks, the absence of interactivity highlights the importance of more visually focused technologies such as Tableau for more dynamic and scalable reporting.

Analysis

Understanding variable connections is critical for making data-driven judgements. For example, correlation analysis might demonstrate how product pricing affects sales success across areas, influencing resource allocation methods. Reflecting on this week's activities, I appreciated how combining statistical tools with multivariate techniques supports effective decision-making and uncovers actionable trends. These skills strengthen my ability to identify patterns and apply insights to professional scenarios, such as optimising supply chain operations.

2.8 Week 8

Description

This week, I concentrated on using Tableau to create interactive visualisations and explore hierarchical data. I worked with features like drill-down capabilities and applied charts as filters to dynamically analyse data across regions. In addition, I looked into Tableau Public for posting dashboards while maintaining data confidentiality and accessibility. Additionally, I explored Tableau Public for publishing dashboards, ensuring data security and accessibility. These activities aligned closely with the module's goal of developing advanced visualisation skills and applying them to practical business scenarios (Solent, 2024).

Feelings

Using Tableau felt intuitive due to my prior experience with Power BI, as both tools share similar functionality. I found the ability to use charts as filters particularly engaging, allowing me to interactively explore regional sales data with ease. While the publishing procedure was fairly straightforward, I had a slight issue changing my Tableau Public password. However, I enjoyed Tableau's privacy reminder, which emphasised the significance of protecting sensitive data before disclosing findings publicly.

Evaluation

Tableau's interactive features showed their effectiveness in simplifying complicated studies. For example, employing hierarchical filters allowed me to dig down from regional patterns to specific cities, emphasising regions that needed attention. These tools provided apparent advantages over static reporting methods, particularly for developing dynamic data narratives. Tableau Public increased the utility of dashboards by making insights available to a wider audience, while the technical difficulty with the password reset revealed potential platform limitations.

Analysis

The ability to create interactive dashboards is crucial for data-driven storytelling in professional settings. For example, tracking product performance using drill-down visualisations can reveal underperforming suppliers, aiding strategic decisions like reallocating resources or discontinuing partnerships. Reflecting on this week's activities, I recognised Tableau's role in making real-time data exploration more accessible and impactful. These skills strengthen my ability to present data insights effectively and adapt visualisations to varying analytical needs.

2.9 Week 9

This week focused on advanced Tableau functionalities, including data blending, field operations like sorting and filtering, and creating calculated fields for custom analyses. These tools allowed for deeper insights by combining multiple datasets, organising data, and generating tailored metrics. Tableau's ability to enhance analytical precision and enable real-time decision-making was a key focus (Milligan, 2019).

Feelings

Data blending stood out as an incredibly useful feature. Combining sales data with customer feedback created a unified view, which revealed how customer satisfaction impacted sales. Unlike manual methods, Tableau's blending tools made this process fast and seamless. I also found calculated fields exciting because they let me create specific metrics, like profit-to-discount ratios, which helped highlight inefficiencies.

Evaluation

Revisiting sorting and filtering techniques helped me understand their potential for real-time analysis. Hierarchical filters made it easy to explore regional sales trends and drill down to underperforming areas. Calculated fields added even more depth, letting me focus on specific business metrics that informed decisions. Tableau's ability to blend datasets and simplify advanced analyses reinforced how valuable it is for business intelligence tasks.

Analysis

This week showed me how Tableau can turn raw data into actionable insights. Data blending helps combine datasets from multiple sources, offering a clearer view of relationships, like how customer satisfaction impacts sales. Tailored metrics created with calculated fields made it easier to identify inefficiencies and refine strategies. Tableau's tools make it a powerful solution for roles that require data-driven decisions.

Action Plan

- Practice data blending with diverse datasets to improve my ability to combine information from multiple sources.
- Explore advanced filtering options to create more interactive and dynamic visualisations.
- Build a library of reusable calculated fields for common business metrics to streamline future work.
- Experiment with data blending scenarios to see how they affect performance and accuracy.

2.10 Week 10

This week focused on creating and publishing interactive dashboards in Tableau. I combined multiple visualisations, added interactive features like filters and highlights, and published insights to Tableau Public. These activities showed how Tableau can make data more engaging and easier for different audiences to explore and understand (Milligan, 2019).

Feelings

I enjoyed creating dashboards in Tableau because the process felt intuitive, especially with my experience using Power BI. A feature I found particularly useful was setting charts as filters, where clicking on a region automatically updated all the data on the dashboard. It made exploring data feel interactive and easy to use.

Publishing to Tableau Public was mostly smooth, though I ran into an issue resetting my password, which was frustrating. That said, I liked Tableau's privacy check before publishing—it's a thoughtful reminder to ensure no sensitive data is shared publicly. It made me appreciate how important it is to be cautious when working with real-world datasets.

Evaluation

Learning how to combine multiple visualisations into a single dashboard was one of the most valuable parts of this week. Features like filters and highlights made the dashboards more dynamic and user-friendly, helping to make data insights accessible to non-technical audiences. Tableau's tools also helped simplify complex data, making it easier to present clear trends and key findings.

Analysis

This week showed me how dashboards can turn data into actionable insights. For example, businesses could use them to track product performance or highlight underperforming suppliers. These insights could lead to smarter decisions, like reallocating resources or improving strategies. Tableau's ability to bring interactivity and storytelling together makes it an essential tool for modern data analysis.

Action Plan

- Build more detailed dashboards with interactive features to improve my skills.
- Try out advanced filtering options to make visualisations more engaging.

- Use data blending to practice combining datasets for better analysis.
- Develop simple design principles for dashboards to make them user-friendly and clear.

3. Conclusion

4. References

Alfred, D. (2024) *Challenges in advanced management accounting*, The Open University. Available at: https://www.open.edu/openlearn/money-business/challenges-advanced-management-accounting/content-section-1 (Accessed: 10 Jan 2025).

IBM (2024) *Data and analytics services on IBM Cloud*. Available at: http://www.ibm.com/analytics/us/en/technology/cloud-data-services/data-scientist/ (Accessed: 10 Jan 2025).

Davenport, T.H. and Patil, D.J. (2012). Data Scientist: The Sexiest Job of the 21st Century. Harvard Business Review.

Marr, B. (2016). Big Data in Practice: How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results. Wiley.

Zachman, J.A. (1987). A Framework for Information Systems Architecture. IBM Systems Journal, 26(3), pp. 276–292.

McKinsey & Company (2020). Analytics Comes of Age: How Organizations Can Scale Advanced Analytics Capabilities. McKinsey & Company.

Bellos, A. (2010). *Alex's Adventures in Numberland: Dispatches from the Wonderful World of Mathematics*. Bloomsbury.

McKinney, W. (2017). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. 2nd ed. O'Reilly Media.

Few, S. (2009). Now You See It: Simple Visualization Techniques for Quantitative Analysis. Analytics Press.

Tableau Software (2021). *The Power of Visual Analytics for Business*. Tableau Whitepaper.

ONS (2023). *Office for National Statistics*. [online] Ons.gov.uk. Available at: https://www.ons.gov.uk/.

Shmueli, G., Patel, N.R., and Bruce, P.C. (2019). Data Mining for Business Analytics: Concepts, Techniques, and Applications in Tableau. Wiley.

Few, S. (2012). Show Me the Numbers: Designing Tables and Graphs to Enlighten. Analytics Press.

Yan, X., & Su, X. (2009). Linear Regression Analysis: Theory and Computing. World Scientific.

Milligan, J. N. (2019). Learning Tableau 2019. Packt Publishing.

Murray, D. (2016). Tableau Your Data!. Wiley-Blackwell.

Agresti, A., & Finlay, P. (2021). Statistical Methods for the Social Sciences (5th ed.). Pearson.

Boud, D., & Feletti, G. (2019). The Challenge of Problem-Based Learning (4th ed.). Routledge.

Field, A. (2018). Discovering Statistics Using IBM SPSS Statistics (5th ed.). Sage Publications.

Solent University. (2024). [Module materials].

Microsoft (2024). *Statistical functions (reference)*. [online] support.microsoft.com. Available at: https://support.microsoft.com/en-us/office/statistical-functions-reference-624dac86-a375-4435-bc25-76d659719ffd.

(Accessed: 10 Jan 2025).

Tableau (2025). *Tableau Training & Tutorials*. [online] Tableau Software. Available at: https://www.tableau.com/learn/training. (Accessed: 07 Jan 2025).

Edmond, S. and Crabtree, M. (2023). *Power BI vs Tableau: Which Should You Choose in 2023?* [online] www.datacamp.com. Available at: https://www.datacamp.com/blog/power-bi-vs-tableau-which-one-should-you-choose. (Accessed: 07 Jan 2025).

Cyntexa. (2024). An Essential Guide to Tableau: Features, Benefits, & How It Works. https://cyntexa.com/blog/essential-guide-to-tableau/ (Accessed: 07 Jan 2025).

5. Appendices A

5.1 Week 1

Who is a Statistician, Data Analyst and Business Analyst?

Who is a Statistician?

A statistician employs statistical methods to collect, assess, and interpret data to solve problems and make informed judgements. Statisticians work at the strategic, managerial, and operational levels of organisations. Their responsibilities include designing experiments, using statistical techniques such as process control, and encouraging methodical methods to decision-making. Furthermore, statisticians play an important role in teaching statistical thinking and ensuring that system implementations are consistent with organisational goals (Abraham, 2007).

Who is a Data Analyst?

A data analyst focuses on processing and interpreting data to derive insights that inform organisational strategies and operations. Their work involves data preparation, statistical analysis, and visualisation to identify patterns and trends. Data analysts use tools like Python, SQL, and Tableau to manage large datasets and present findings in actionable formats. Industry expectations highlight a combination of technical proficiency, domain expertise, and communication skills as critical to their success (Data Analyst Competencies, n.d.).

Who is a Business Analyst?

A business analyst focuses on closing the gap between business requirements and technological solutions. They are responsible for acquiring, evaluating, and validating requirements to drive changes in processes, systems, and policies. Interviews, workshops, requirement documentation, and process model creation are all important responsibilities. corporate analysts ensure that solutions meet corporate objectives and successfully solve organisational challenges. Their tasks include facilitating communication and translating corporate needs into actionable solutions (Brandenburg, 2021).

Using the Web and online tools, identify the necessary skills required for becoming a data scientist.

How can you position yourself for any of these job roles?

To succeed as a statistician, data analyst, business analyst, or data scientist, it is important to build the right mix of skills, knowledge, and practical experience, tailored to each role.

Statistician:

Statisticians focus on applying mathematics and statistics to solve problems and make decisions. Learning tools are R, SAS, or SPSS and more. They work across all levels of an organisation strategic, managerial, and operational—using statistical thinking to improve processes and decision-making (Abraham, 2007).

Data Analyst:

Data analysts use tools like Python, SQL, Tableau, and Power BI to turn raw data into insights. They need to analyse data, create dashboards, and explain their findings clearly. Building a portfolio with real-world projects helps to show your abilities in this competitive field (Data Analyst Competencies, n.d.).

Business Analyst:

Business analysts connect business needs with technical solutions. They gather and analyse requirements using tools like UML and BPMN, and they often work in Agile environments. Communication and problem-solving are key, and certifications like CBAP can make you stand out. Experience working with stakeholders helps align solutions with business goals (Brandenburg, 2021).

Data Scientist:

Data scientists focus on finding patterns and insights in large datasets using programming, machine learning, and tools like Hadoop and Spark. They need strong skills in Python and statistics, along with practical experience from projects or internships. Advanced education like a master's degree can also help in this role (Smaldone et al., 2022).

State and discuss the requirements of an organisation before hiring a data scientist, for example, access or availability of large amount of data.

Requirements	Description
Data Availability and Infrastructure	Access to large, diverse datasets and robust tools for data processing, storage, and analysis.
Clearly Defined Business Objectives	Establish clear goals, such as improving retention, optimising operations, or exploring new markets.
Analytical Culture	Promote a data-driven culture with strong data literacy and support for data-informed decision-making.
Investment in Tools and Technologies	Provide tools such as Python, R, machine learning libraries, and visualisation platforms for analysis.
Collaboration and Teamwork	Encourage communication and collaboration with software engineers, business analysts, and domain experts.
Commitment to Talent Development	Offer opportunities for skill development, such as certifications, workshops, and conferences.

Sources:

Smaldone, M., et al. (2022). *Employability Skills: Profiling Data Scientists in Industry*. European Management Journal.

What is Business Intelligence BI?

Business Intelligence (BI) refers to a collection of tools, procedures, and methods that turn unprocessed data into insights that may be used. Businesses may monitor performance, analyse data, and assist in strategic decision-making with the help of BI (Chen et al., 2012).

Of what benefits is business intelligence to any organisation?

Benefit	Description
Improved Decision-Making	Provides accurate, timely insights for data-driven decisions and supports predictive analysis and trend identification.
Operational Efficiency	Identifies bottlenecks, optimizes workflows, and tracks performance metrics to streamline processes.
Customer Insights	Analyses customer behaviours and preferences, enhancing personalisation and customer satisfaction.
Competitive Advantage	Anticipates market trends, adapts strategies proactively, and strengthens market positioning with predictive tools.
Resource Optimisation	Improves financial management, supply chain performance, and enables better planning through data analysis.

Sources:

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, 36(4), 1165-1188.

State at least 3 job roles available in the areas of business intelligence indicating the skills, starting salary, maximum salary and other job benefits.

Data Scientist

Skills required:

- Advanced statistical analysis
- Machine learning
- Programming (Python, R)
- Big data technologies
- Data mining
- Predictive modeling
- Data visualisation

Salary range:

- Starting salary: Around £50,000
- Maximum salary: Can exceed £100,000 for experienced professionals

Additional benefits:

- Flexible working hours
- Opportunities for continuous learning and skill development

Collaboration with cross-functional teams

BI Solutions Architect

Skills required:

- Data warehouse design
- ETL processes
- BI tool expertise (e.g., Tableau, Power BI)
- Cloud computing knowledge
- Project management
- System integration
- Performance optimisation

Salary range:

- Starting salary: Approximately £70,000
- Maximum salary: Can exceed £120,000 for senior positions

Additional benefits:

- Remote work options
- Leadership opportunities
- Involvement in strategic decision-making

Potential for bonuses based on project success

Business Intelligence Analyst

Skills required:

- Data analysis
- SQL programming
- Data visualisation
- Business acumen
- Problem-solving
- Data warehousing
- Report generation
- Predictive analytics

Salary range:

- Starting salary: Approximately £33,183
- Maximum salary: Can exceed £45,073 for senior positions

Additional benefits:

- Free medical and dental health cover
- Opportunity for salary increase after one year of training
- Access to sensitive intelligence material (for certain positions)

Sources:

Coursera. (2025, January 6). Data scientist salary guide: What to expect in 2025. https://www.coursera.org/gb/articles/data-scientist-salary

Coursera. (2025, January 8). Business intelligence analyst salary guide. https://www.coursera.org/articles/business-intelligence-analyst-salaryProspects. (n.d.).

Data scientist job profile. Retrieved January 12, 2025, from https://www.prospects.ac.uk/job-profiles/data-scientist

ProjectPro. (2024, October 28). Data scientist salary - The ultimate guide for 2024. https://www.projectpro.io/article/data-scientist-salary-the-ultimate-guide-for-2021/218

Userpilot. (2024, September 25). What is a business intelligence analyst? Responsibilities, salaries. https://userpilot.com/blog/what-is-a-business-intelligence-analyst/

Userpilot. (2024, September 25). Business intelligence analyst salary [+ Resources to advance]. https://userpilot.com/blog/business-intelligence-analyst-salary/

- Compare and contrast the characteristics of the identified job roles and present a statement of conclusion.

Business Intelligence Analysts, Data Scientists, and BI Solutions Architects play distinct but interconnected roles within organisations, each contributing to the effective use of data. BI Analysts focus on historical data analysis and visualisation to support decision-making (Coursera, 2025a; Userpilot, 2024a). Data Scientists employ advanced analytics, machine learning, and big data technologies to uncover insights and predict future trends (Coursera, 2025b;

ProjectPro, 2024; Prospects, n.d.). On the other hand, BI Solutions Architects take a broader systems-oriented approach, designing and implementing the infrastructure that enables the other two roles to function effectively (Userpilot, 2024b). Together, these roles highlight the spectrum of skills and expertise required to drive data-driven strategies and innovation in organisations.

5.2 Week 2

Dataset: Sample-Superstore

The Sample - Superstore dataset covers transactional data such as sales, earnings, customer groups, and product categories. This dataset is commonly used for business analytics and data visualisation projects.

Objective:
Summary statistics (mean, median, mode, standard deviation, range, and mean deviation).
Identifying and analysing outliers
Analytical findings and insights are interpreted.
Any suggestions or observations based on the data

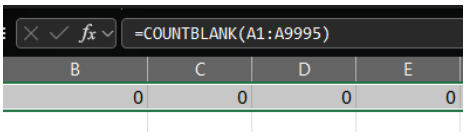


Figure 1 Blank Cell Count Formula

The COUNTBLANK function in Excel confirmed that the dataset contained no missing values, maintaining the integrity of subsequent analyses.

Statistic	Values	Formulas
Mean	229.8580008	(=AVERAGE(R1:R9995))
Mean Deviation	268.1563131	
Standard Deviation	623.2139188	(=STDEV.P(R1:R9995))
Mode	12.96	(=MODE.SNGL(R1:R9995))
Minimum	0.444	(=MIN(R2:R9995))
First Quartile	17.28	(=QUARTILE.INC(R2:R9995, 1))
Median	54.49	(=MEDIAN(R2:R9995))
Third Quartile	209.94	(=QUARTILE.INC(R2:R9995, 3))
Maximum	22638.48	(=MAX(R2:R9995))
Interquartile Range (IQR)	192.66	(=QUARTILE.INC(R2:R9995, 3)-QUARTILE.INC(R2:R9995, 1))
Lower Bound	-271.71	(=17.28-(1.5*192.66))
Upper Bound	498.93	(=209.94+(1.5*192.66))

Figure 2 Sales Summary Statistics

The summary statistics calculated with Excel functions highlight the essential metrics for the sales data, such as mean, median, mode, standard deviation, and IQR. The mean sales value (229.86) is much higher than the median value (54.49), indicating a positively skewed distribution. This skew indicates that a small fraction of high-value sales inflates

the mean, rendering the median a more trustworthy measure of central tendency in this dataset (Kaur, Stoltzfus, and Yellapu, 2018). The standard deviation (623.21) indicates significant fluctuation in sales numbers. According to descriptive statistical studies (Kaur, Stoltzfus, and Yellapu, 2018), this heterogeneity could be attributed to product category diversity, bulk orders, and applied discounts.

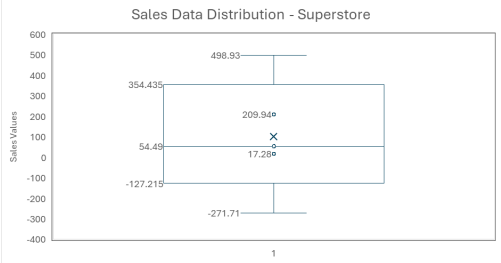


Figure 3 Sales Data Distribution

The box plot illustrates the dispersion and variability of sales numbers in the dataset. The upper bound for identifying outliers was calculated as 498.93 using the formula:

Upper Bound=Q3+(1.5×IQR)

Sales values exceeding this threshold are considered outliers, likely caused by large infrequent purchases. As noted in Kaur, Stoltzfus, and Yellapu (2018), examining these outliers is crucial to understanding their validity and assessing their influence on overall sales trends.

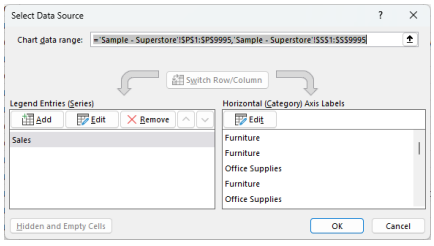


Figure 4 Chart Data Source Setup

This figure shows how to set up Excel to create category-based visualisations. It shows how data sources are defined for charts by defining sales numbers and product categories. This configuration is critical for constructing relevant visual studies, such as bar charts and scatter plots, to investigate patterns across multiple categories.

Figure 5 Sales Distribution by Product Category

This scatter plot highlights the distribution of sales values for three product categories: furniture, office supplies, and technology. The Technology category has the greatest variation in sales, including big outliers. The presence of extreme outliers, as shown in the distribution by product category, amplifies variability in the dataset.

A	B	C	D
Row ID	Order ID	Order Date	
1	CA-2016-152156	08/11/2016	
2	CA-2016-152156	08/11/2016	
3	CA-2016-138688	12/06/2016	
4	US-2015-108966	11/10/2015	
5	US-2015-108966	11/10/2015	
6	CA-2014-115812	09/06/2014	
7	CA-2014-115812	09/06/2014	
8	CA-2014-115812	09/06/2014	
9	CA-2014-115812	09/06/2014	
10	CA-2014-115812	09/06/2014	

Figure 6 Adding new Column

```
=TEXT(C2,"YYYY-MM")
```

C	D
Order Date	Order Month
08/11/2016	2016-11

Figure 7 Order Month Formula

This figure 6 and 7 demonstrates the creation of a new column, Order Month, derived from the Order Date column using the formula =TEXT(C2, "YYYY-MM"). This additional column facilitates temporal analysis of sales trends. By categorising data by months, it becomes possible to identify seasonal patterns or trends, such as peak sales periods or

recurring dips, enabling data-driven decision-making.

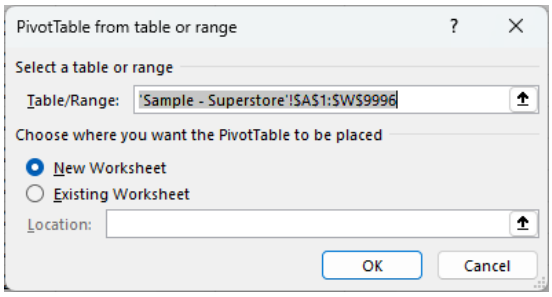


Figure 8 Pivot Table Setup

This figure illustrates the setup of a Pivot Table in Excel, enabling dynamic data summarisation for the Sample - Superstore dataset. The selected data range (Sample - Superstore!$A$1:$W$9999) encompasses all relevant fields necessary for comprehensive analysis. The Pivot Table provides a robust tool for aggregating sales and profit data across multiple dimensions, such as product categories, customer segments, and time periods. This setup lays the foundation for generating actionable insights by summarising large volumes of data efficiently and flexibly.

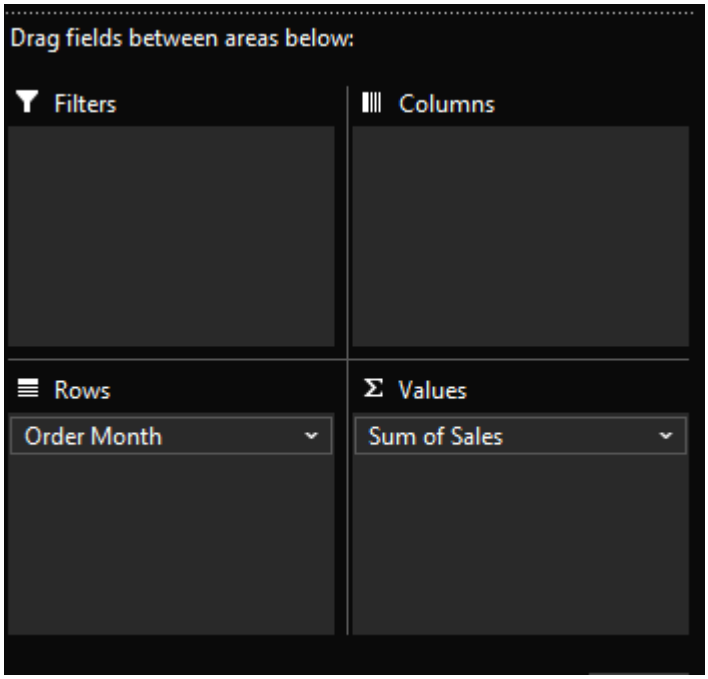


Figure 9 Pivot Table Fields Setup

This figure displays the configuration of fields in the Pivot Table used for the analysis. The Order Month column is in the Rows part to organise sales data by month, the Sum of Sales field is in the Values section to calculate total sales for each month. This approach allows for straightforward monthly trend analysis, which is critical for evaluating temporal sales performance.

Row Labels	Sum of Sales
2014-01	14236.895
2014-02	4519.892
2014-03	55691.009
2014-04	28295.345
2014-05	23648.287
2014-06	34595.1276
2014-07	33946.393
2014-08	27909.4685
2014-09	81777.3508
2014-10	31453.393
2014-11	78628.7167
2014-12	69545.6205

Figure 10 Monthly Sales Summary

This table showcases the monthly total sales figures for 2014, generated via a Pivot Table. The data highlights the variability in sales, with a significant peak in September (81,777.35) and the lowest sales in February (4,519.89). This variability suggests potential seasonal trends or periodic factors influencing sales performance.

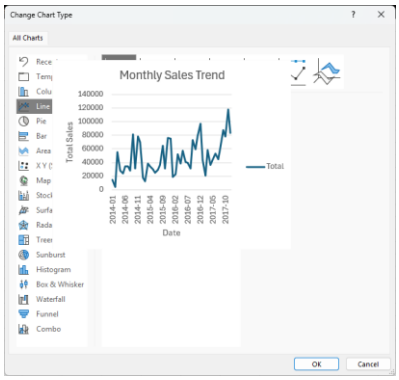


Figure 11 Line chart Setup

This figure illustrates the setup of a line chart used to visualise the monthly sales trend. The chart type was configured to display the Total Sales across the timeline, providing an intuitive representation of sales fluctuations and patterns over months. Such visualisations are pivotal for identifying trends, peaks, and dips in sales, which can guide decision-making and strategic planning.

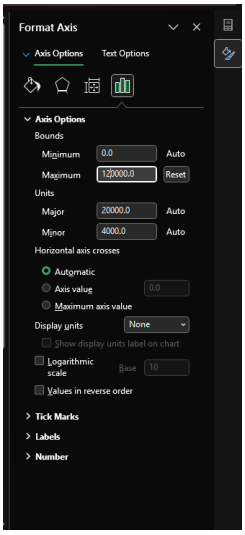


Figure 12 Axis Formatting Options

This figure shows the configuration settings for formatting the axes of the Monthly Sales Trend line chart. Adjusting the Bounds and Units enhances the clarity of the visualisation by ensuring appropriate scaling of the data points. These formatting options are crucial for accurately displaying trends, making patterns more discernible, and avoiding distortion in the representation of the dataset.

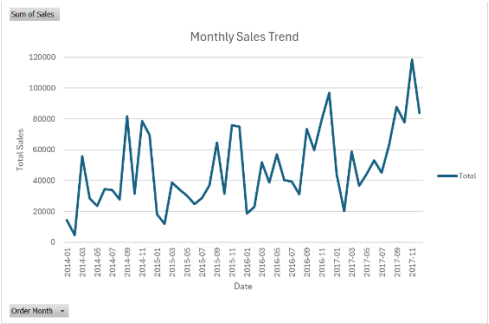


Figure 13 Monthly Sales Trend

This figure shows the configuration settings for formatting the axes of the Monthly Sales Trend line chart. Adjusting the Bounds and Units enhances the clarity of the visualisation by ensuring appropriate scaling of the data points. These formatting options are crucial for accurately displaying trends, making patterns more discernible, and avoiding distortion in the representation of the dataset.

	A	B	C
3	**Row Labels**	**Sum of Sales**	**Sum of Profit**
4	Furniture	741999.7953	18451.2728
5	Office Supplies	719047.032	122490.8008
6	Technology	836154.033	145454.9481
7	**Grand Total**	**2297200.86**	**286397.0217**

Figure 14 Category-Wise Sales and Profit Summary

This pivot table summarises the total sales and profit figures for each product category: Furniture, Office Supplies, and Technology. The "Grand Total" row provides the overall figures for both metrics across all categories. This summary facilitates a deeper understanding of the relative performance and profitability of each category, enabling targeted business strategies.

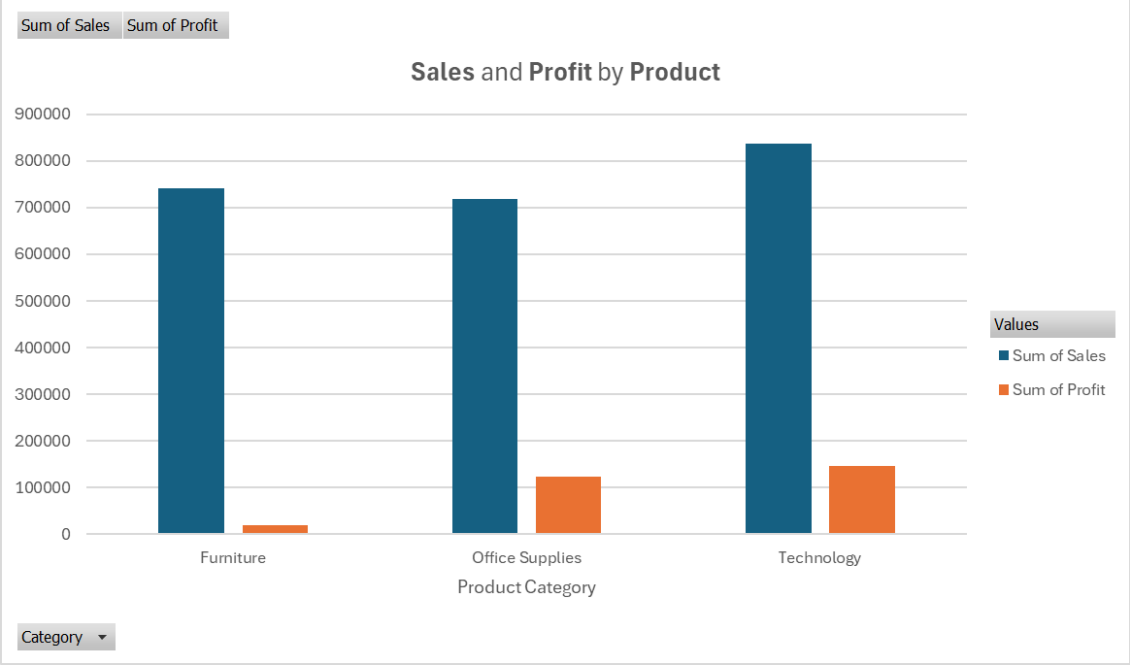


Figure 15 Sales and Profit by Product

This bar chart visually compares the total sales and profit for each product category (Furniture, Office Supplies, Technology). It highlights that while Technology generates the highest sales and profit, Furniture has a significantly lower profit margin relative to its sales figures. This insight suggests potential inefficiencies or higher costs associated with the Furniture category.

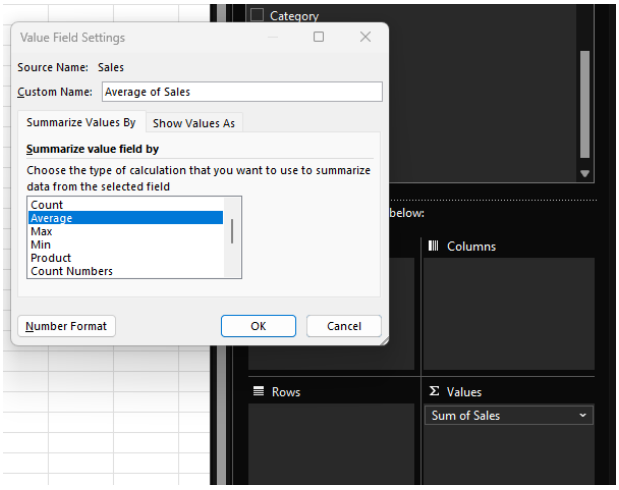


Figure 16 Value Field Settings Configuration

This figure shows the configuration used to calculate the "Average of Sales" within the pivot table. The settings illustrate how Excel's pivot table functionality is leveraged to summarise data accurately, enabling deeper analysis such as identifying average sales trends across categories or segments.

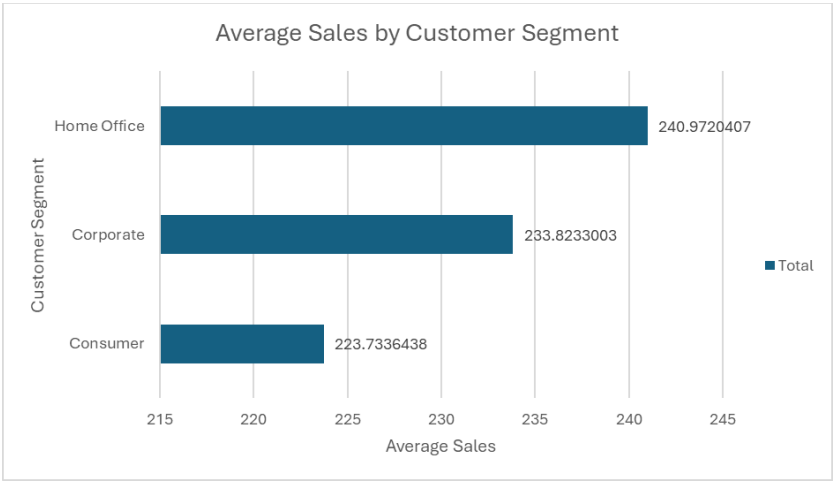


Figure 17 Average Sales by Customer Segment

This bar chart highlights the average sales values across different customer segments: Home Office, Corporate, and Consumer. Home Office customers have the highest average sales value, followed closely by the Corporate segment, with Consumers having the lowest average sales. This differentiation can guide targeted marketing strategies and resource allocation for each segment to maximise revenue opportunities.

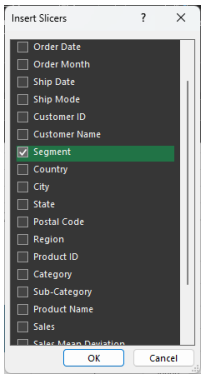


Figure 18 Insert Slicer Options

The slicer configuration screen demonstrates the ability to filter data dynamically within pivot tables and charts by selecting specific attributes such as customer segments, product categories, or regions. Slicers improve interactivity and provide users with flexible insights tailored to specific criteria.

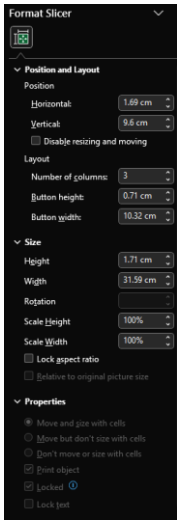


Figure 19 Slicer Formatting Options

This screenshot demonstrates the slicer formatting settings used to customise the appearance and positioning of slicers within the dashboard. Adjustments include the number of columns, button height, button width, and overall slicer size. These settings ensure that the slicers are visually aligned with the dashboard layout, enhancing usability and presentation quality.

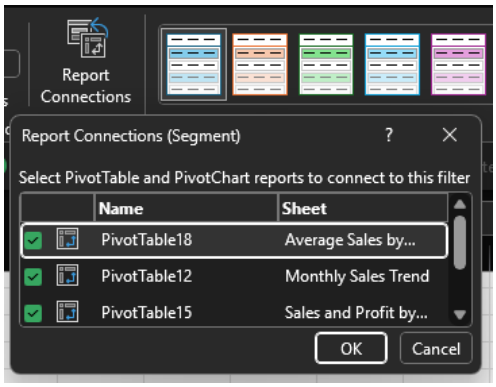


Figure 20 Report Connections

This figure highlights the process of linking slicers to multiple PivotTables or PivotCharts. By enabling report connections, the slicers synchronise filters across related visualisations, ensuring consistency and improving the interactivity of the dashboard. For instance, in this case, the "Segment" slicer is connected to three key elements: Average Sales by Customer Segment, Monthly Sales Trend, and Sales and Profit by Product.

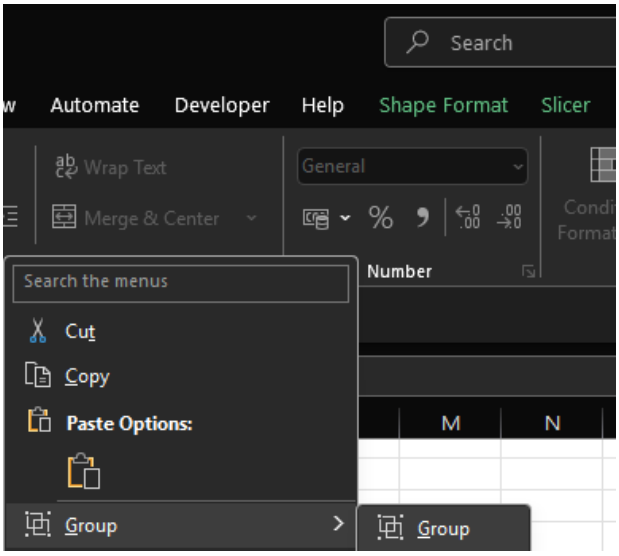


Figure 21 Group Context Menu

This figure demonstrates the use of the Group Context Menu for organising visual elements. Grouping allows multiple slicers or charts to be moved or formatted collectively, ensuring alignment and enhancing the overall layout of the dashboard.

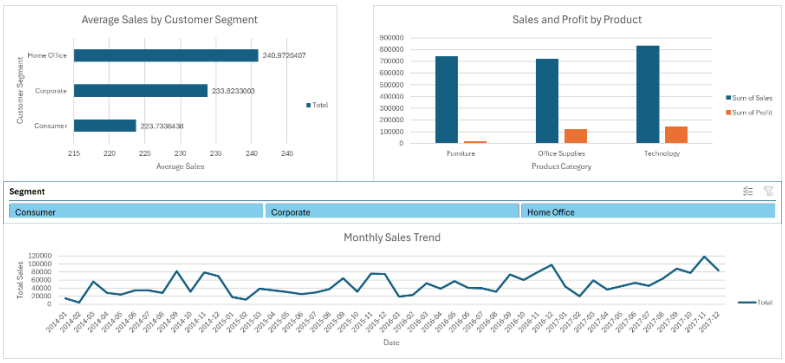


Figure 22 Filter applied to All segments

This figure showcases the dashboard with the filter set to include all customer segments: Consumer, Corporate, and Home Office. The data and visualisations update dynamically to reflect the contributions of each segment, providing a holistic view of the sales trends, product profitability, and customer segment analysis.

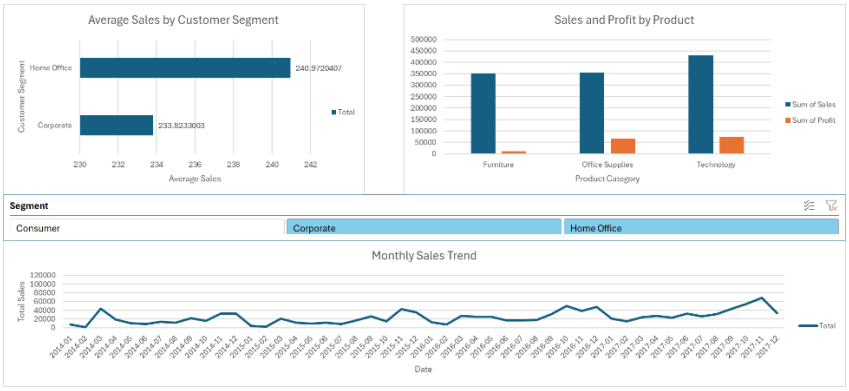


Figure 23 Filter excluding Consumer

This figure demonstrates the impact of excluding the Consumer segment using the slicer. The resulting changes in the Average Sales by Customer Segment, Sales and Profit by Product, and Monthly Sales Trend visualisations highlight the influence of the Corporate and Home Office segments on sales and profitability.

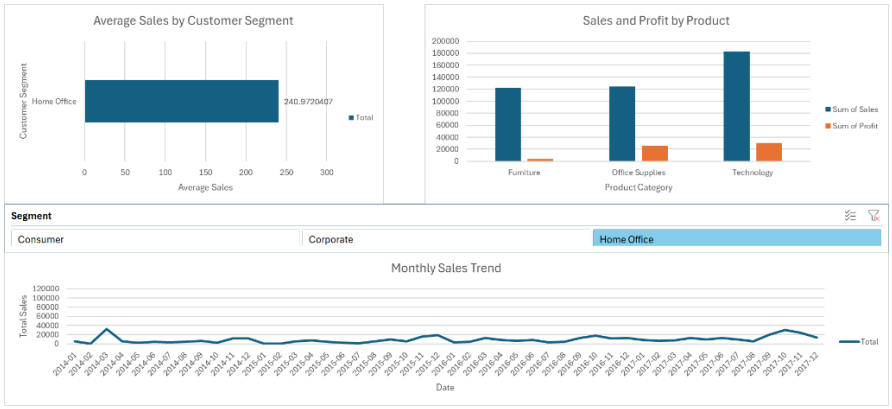


Figure 24 Filter applied to Home Office

This figure highlights the dashboard filtered exclusively for the Home Office segment. The visualisations, including Average Sales by Customer Segment, Sales and Profit by Product, and Monthly Sales Trend, dynamically adapt to display data specific to this segment. This approach enables a focused analysis of the sales patterns, product performance, and revenue trends associated with Home Office customers, aiding in segment-specific decision-making and strategy development.

Summary:

The analysis of the Sample-Superstore dataset highlights actionable pathways for improving business performance. Recommendations emphasise the strategic enhancement of profitable areas and rectifying inefficiencies. Given the critical role of Technology as the most profitable category, businesses should expand offerings in this area, particularly targeting Home Office and Corporate segments. These customer groups exhibit the highest average sales, reflecting their significance in driving revenue. As discussed by Kaur, Stoltzfus, and Yellapu (2018), understanding central tendency metrics like the mean and median is vital for tailoring business strategies. Here, the positive skew of sales data indicates that a minority of high-value transactions disproportionately impact revenue, requiring closer attention to these lucrative customer behaviours.

The Furniture category demands immediate evaluation. Despite substantial sales, its disproportionately low profit margins indicate operational inefficiencies, potentially in areas such as supply chain logistics or pricing models. Correcting such inefficiencies aligns with the concept of identifying variance in data to pinpoint performance gaps, as Kaur et al. (2018) advocate. Seasonal trends also warrant attention, with significant Q4 sales peaks suggesting the need for refined inventory management and marketing strategies. Aligning these efforts with predicted high-demand

periods ensures resource efficiency and customer satisfaction.

Conclusively, this analysis demonstrates the value of dynamic visualisation tools and statistical methods in guiding data-driven decisions. By applying descriptive statistical insights, such as those emphasised by Kaur, Stoltzfus, and Yellapu (2018), businesses can optimise operations, enhance profitability, and adapt strategies to customer behaviours and market dynamics. These findings provide a foundation for continuous improvement and future analytics, paving the way for sustainable growth in a competitive environment.

5.3 Week 3

Variable	Type of Variable	Justification
Row ID	Discrete	Unique identifier for each row, typically a numerical count, non-continuous.
Order ID	Nominal	Categorically distinguishes each order, used as an identifier not implying mathematical operations.
Order Date	Ordinal	Dates have a natural order but the interval between them is not consistent across the dataset.
Order Month	Ordinal	Similar to Order Date, months have a natural order without consistent intervals, suitable for time trends.
Ship Date	Ordinal	Follows an order but does not represent a consistent interval, useful for lead time calculations.
Ship Mode	Nominal	Categorizes shipping methods without any intrinsic ordering.
Customer ID	Nominal	Unique identifier for each customer, used for categorization without numerical operations.
Customer Name	Nominal	Names serve as labels without quantitative value or order.
Segment	Nominal	Represents market segments, which are categorical labels without quantitative value.
Country	Nominal	Used to categorize data by country without any quantitative implication.
City	Nominal	Locational data used for categorization, without implying any order or interval.
State	Nominal	States function as categorical labels in analysis, similar to City.
Postal Code	Nominal	Used for geographical sorting and categorization, numerical but not used for calculations.
Region	Nominal	Categorical area designation without inherent numerical value or order.
Product ID	Nominal	Unique identifier for products, used for tracking and categorization.
Category	Nominal	General product categories that are used to classify products without implying order.
Sub-Category	Nominal	More specific product classifications within a category, no intrinsic order.
Product Name	Nominal	Names are categorical and used for identification without quantitative measurement.
Sales	Continuous	Quantitative measurement of revenue, which can have any value within a range and involve calculations.
Quantity	Discrete	Counts of items per transaction, involves discrete numerical values and simple arithmetic.
Discount	Continuous	Represents a rate, which can vary continuously across a range and impact sales calculations.
Profit	Continuous	Financial metric calculated from sales and costs, continuous as it can take any value within a range.
Sales Mean Deviation	Continuous	Represents a statistical measure of variance from the mean, can be any value within a range.

Figure 25 Variable Types Table

Understanding Variable Types and Descriptive Statistics

Before performing any statistical analysis, it is important to understand the types of variables in the dataset. This helps select appropriate descriptive statistics (Kaur, Stoltzfus, and Yellapu, 2018). For example, measures of central tendency summarize data, but different measures suit different variable types. The mean is suitable for continuous, normally distributed data, like Sales and Profit. The median, however, is better for ordinal data, such as Order Month, or continuous data that's not normally distributed (Kaur, Stoltzfus, and Yellapu, 2018). Measures of dispersion, such as range, variance, and standard deviation, are typically applied to continuous data.

Classification of Descriptive Statistics

Descriptive statistics can be further classified as:

- Measures of Frequency: These show how often a value occurs and are used for all variable types, but are especially useful for categorical variables, such as Ship Mode, Segment, Category (Kaur, Stoltzfus, and Yellapu, 2018). Examples include frequency, ratios, rates, proportions, and percentages.
- Measures of Position: These describe where values fall in relation to each other and are used for ordinal, interval, and ratio data. They include percentile ranks and quartile ranks.

Applying Appropriate Descriptive Statistics

Using the correct descriptive statistic for each variable type avoids misleading results and allows meaningful conclusions to be drawn from the data (Kaur, Stoltzfus, and Yellapu, 2018).

1. Which product category has the highest average profit?

Using the =AVERAGEIF() function:

Average Profit by Product Category		
Category	Avg. Profit ($)	Formula
Technology	78.75200222	(=AVERAGEIF($P$2:$P$9995, "Technology", $V$2:$V$9995))
Office Supplies	20.32704959	(=AVERAGEIF($P$2:$P$9995, "Office Supplies", $V$2:$V$9995))
Furniture	8.69932711	(=AVERAGEIF($P$2:$P$9995, "Furniture", $V$2:$V$9995))

Figure 26 Average Profit by Product Category Using AVERAGEIF Function

The Product Category with the highest Average Profit is Technology with $78.75

2. Which customer segment has the highest total sales?

Using the SUMIF function:

Toltal Sales by Segment		
Segment	Total Sales ($)	Formula
Consumer	1161401.345	(=SUMIF($I$2:$I$9995, "Consumer", $S$2:$S$9995))
Corporate	706146.3668	(=SUMIF($I$2:$I$9995, "Corporate", $S$2:$S$9995))
Home Office	429653.1485	(=SUMIF($I$2:$I$9995, "Home Office", $S$2:$S$9995))

Figure 27 Total Sales by Segment Using SUMIF Function

The segment with the highest total sales is Consumer with $1,161,401.05

3. What is the highest sales value for each product category?

Using the MAXIFS function:

Highest Sale per Product Category		
Category	Highest Sale ($)	Formula
Technology	22638.48	(=MAXIFS($S$2:$S$9995, $P$2:$P$9995, "Technology"))
Office Supplies	9892.74	(=MAXIFS($S$2:$S$9995, $P$2:$P$9995, "Office Supplies"))
Furniture	4416.174	(=MAXIFS($S$2:$S$9995, $P$2:$P$9995, "Furniture"))

Figure 28 Highest Sale per Product Category Using MAXIF Function

The highest sale value is $22638.48 for Technology, $9892.74 for Office Supplies and $4416.174 for Furniture being the lowest between the three.

Comparison

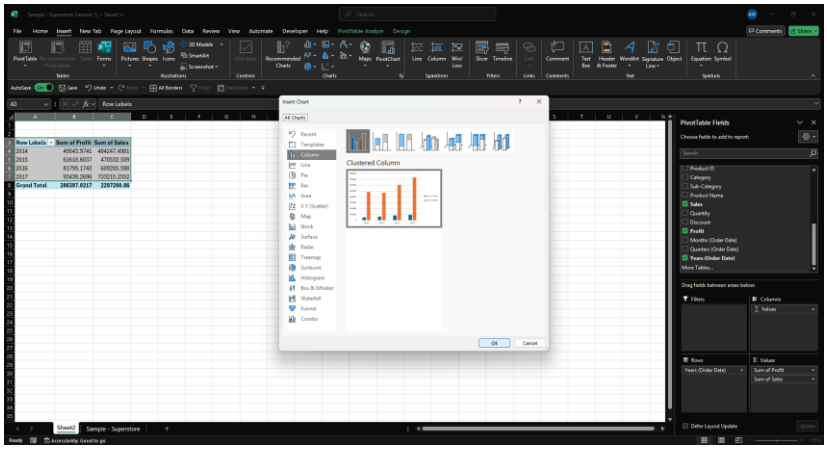


Figure 29 Setup of chart

Figure 30 Sales and Profit by Year

The chart compares the Sum of Sales and Sum of Profit across different years, categorising it as a comparison chart. The clustered column format effectively highlights the differences between these two metrics over time. It demonstrates that both Sales and Profit increase substantially year by year, although on different scales.

Dataset Overview

The dataset contains transactional data with 23 variables, including identifiers (Row ID, Order ID, Customer ID, Product ID), temporal attributes (Order Date, Ship Date, Order Month), categorical variables (Ship Mode, Segment, Region, Category, Sub-Category), and quantitative metrics. These variables provide insights into sales success, customer behaviour, and product profitability across locations and time periods. Categorical variables provide segmentation and comparison, whereas quantitative variables allow for detailed statistical and trend analysis.

Observed Trends

Analysis of the dataset reveals several noteworthy trends:

Sales and profit growth have been consistent over the years, with large increases in both categories. However, sales growth far outpaces profit growth, indicating possible opportunities for cost-cutting or increased efficiency.

Category-Wide Profitability: Technology has the highest average profit per sale, followed by Office Supplies and Furniture. This shows the profitability of the Technology category and may suggest prioritising marketing or inventory allocation.

Regional and Segment Variations: Preliminary research suggests possible variances in performance across customer segments and geographies, highlighting the possibility of customised strategies based on consumer preferences and demand trends.

5.4 Week 4 NOT DONE

When assessing the quality of my dataset, I looked at several key dimensions to make sure the data was suitable for analysis:

Accuracy:
All the values in the dataset seem accurate. Grades are between 51 and 100, study hours range from 1 to 5, and attendance rates are percentages between 60.22% and 99.48%. Nothing looks out of place.

Completeness:
There are no missing values in the dataset. Every column—Student_ID, Grade, Study_Hours, and Attendance_Rate—has data for all 33 students.

Consistency:
The dataset is consistent. Study hours are all integers, grades are numeric, and attendance rates are formatted as percentages. Everything is as it should be.

Validity:
All the data is valid and fits the expected formats. Attendance rates are within a realistic range (0–100), grades are reasonable (0–100), and study hours are whole numbers as expected.

Integrity:
The dataset has no duplicate student IDs, so each student's information is unique. This ensures there are no errors caused by repeating data.

The dataset is complete, accurate, consistent, and valid. There were no issues with missing or incorrect data, so it is ready for analysis without needing further adjustments.

Figure 33 Task 4

Five Questions to Explore

1. **What is the average grade for students who study 4 or more hours?**

For this question, I looked at how studying for at least 4 hours affects grades. I used Excel to filter the dataset, so it only included students who studied 4 or more hours. Then, I calculated the average grade for these students using the =AVERAGE function. The result came out as 79.75.

This shows that students who spend more time studying tend to get better grades compared to the class average. It's clear that longer study hours are helpful, but it's also possible that other factors like attendance or personal study habits play a role in performance. Encouraging students to study at least 4 hours could improve their grades, but additional support, like teaching study skills, might help even more.

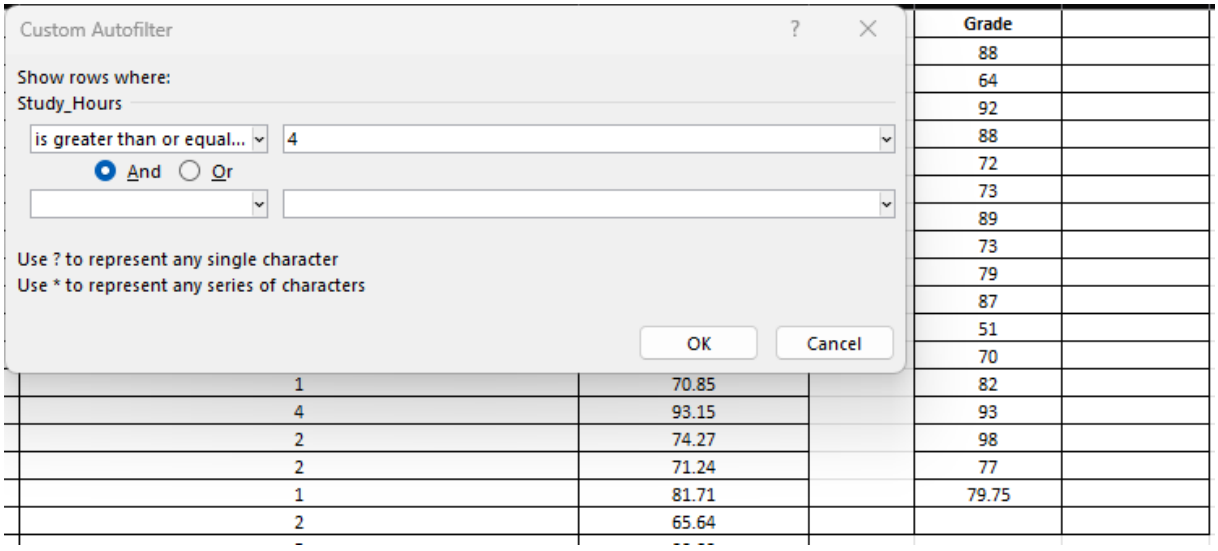


Figure 34 Task 4

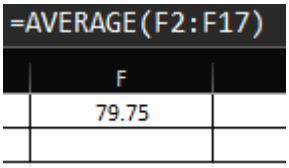


Figure 35 Task 4

Grade
88
64
92
88
72
73
89
73
79
87
51
70
82
93
98
77

Figure 36 Task 4

2. **What is the most common grade range in the class?**

To find the most common grade range, I grouped the grades into ranges like 50–60, 61–70, and so on. Using Excel, I counted how many students fell into each range by applying the =COUNTIFS() formula. The ranges and their counts were:

50–60: 6 students =COUNTIFS(B2:B34, ">=51", B2:B34, "<=60")

61–70: 5 students =COUNTIFS(B2:B34, ">=61", B2:B34, "<=70")

71–80: 11 students =COUNTIFS(B2:B34, ">=71", B2:B34, "<=80")

The range with the highest count was **71–80**, with 11 students achieving grades in this range.

This shows that the majority of students in the class perform within the 71–80 range, which is slightly above average. It indicates a concentration of students in the middle performance band, suggesting that while many are doing reasonably well, there is still room for improvement to push more students into higher grade ranges.

Figure 37 Task 4

Question 3: Do Attendance and Study Hours Correlate with Grades?

To explore the relationship between grades, attendance, and study hours, I used Excel's =CORREL function to calculate two correlations:

Attendance Rate and Grades:

The correlation between attendance and grades is 0.30. This indicates a weak positive relationship, meaning that higher attendance is somewhat associated with better grades, but it is not a strong predictor.

Study Hours and Grades:

The correlation between study hours and grades is 0.33, which also shows a weak positive relationship. This suggests that while students who study more tend to have higher grades, the link is not particularly strong.

These results highlight that neither attendance nor study hours alone have a strong impact on academic performance. Other factors, such as the quality of study or individual aptitude, might play a larger role in influencing grades. Encouraging students to both attend class regularly and study consistently could still have a positive impact overall.

Figure 38 Task 4

(=CORREL(D2:D34, B2:B34))	0.303228612
(=CORREL(C2:C34, B2:B34))	0.330913156

Figure 39 Task 4

Question 4: What is the average attendance rate for students scoring above 80?

From the analysis, I filtered the dataset to only include students who scored above 80. Using Excel's filter tool and the formula =AVERAGE(range), I calculated the average attendance rate for these students. The result is **84.73%**.
This shows that students scoring higher than 80 generally have high attendance rates. It highlights that consistent attendance is likely a key factor contributing to better performance. However, attendance alone might not explain all results, as other aspects like study habits or academic ability could also play a role.

Figure 40 Task 4

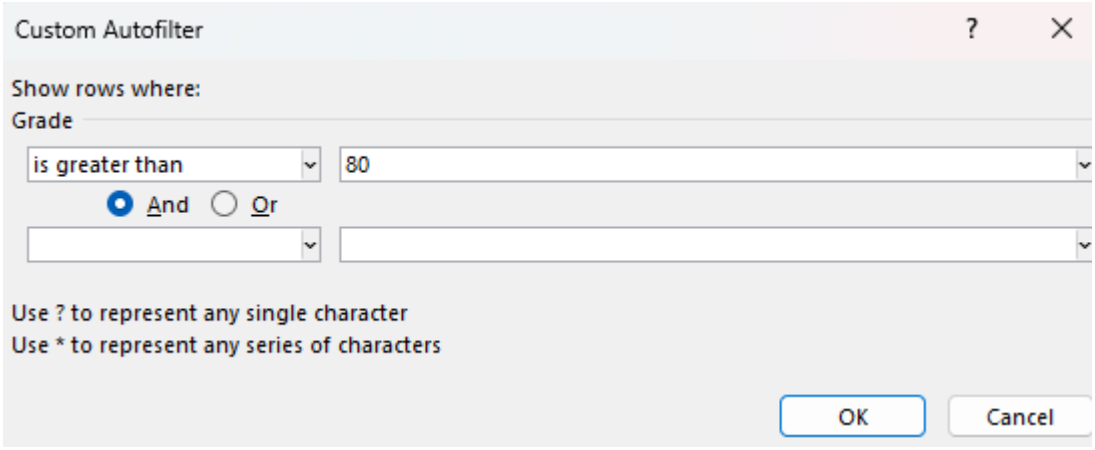


Figure 41 Task 4

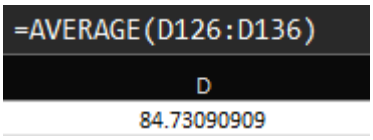


Figure 42 Task 4

	Attendance_Rate for Grade < 80
	67.39
	97.58
	96.87
	70.85
	93.15
	62.98
	90.89
	92.62
	90.85
	74.34
	94.52
AVG	84.73090909

Figure 43 Task 4

Question 5: What is the relationship between Study Hours and Grades?

To explore how study hours relate to grades, I created a scatter plot in Excel with Study Hours on the X-axis and Grades on the Y-axis. I added a trendline, displayed the equation, and calculated the R-squared value.

The trendline equation is y = 3.28x + 64.71, which shows that grades increase by around 3.28 points for every extra hour of study. However, the R-squared value is **0.1095**, meaning the relationship is weak. This suggests that while studying more is generally beneficial, other factors like attendance, learning habits, or motivation likely play a bigger role in determining grades.

This analysis shows a weak positive trend between study hours and grades, but it is clear that study hours alone do not fully explain student performance.

Figure 44 Task 4

```
=CORREL(C2:C34, B2:B34)
```

C
0.330913156

Figure 45 Task 4

Grade	Study_Hours
88	4
78	3
64	4
92	4
57	1
70	3
88	5
68	3
72	5
60	1
60	2
73	4
85	1
89	4
73	2
52	2
71	1
51	2
73	5
93	2
79	4
87	4
51	4
70	4
82	5
61	3
71	1
93	4
74	2
98	4
76	2
91	2
77	4

Figure 46 Task 4

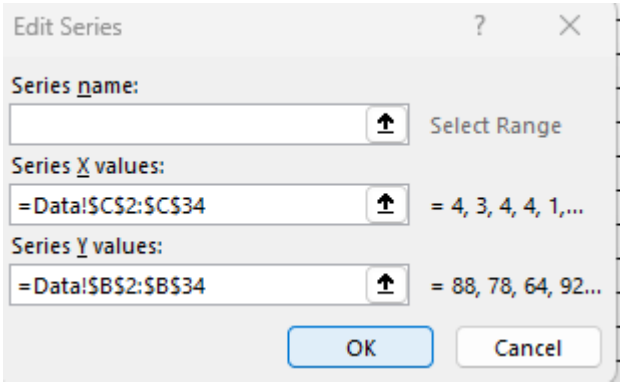


Figure 47 Task 4

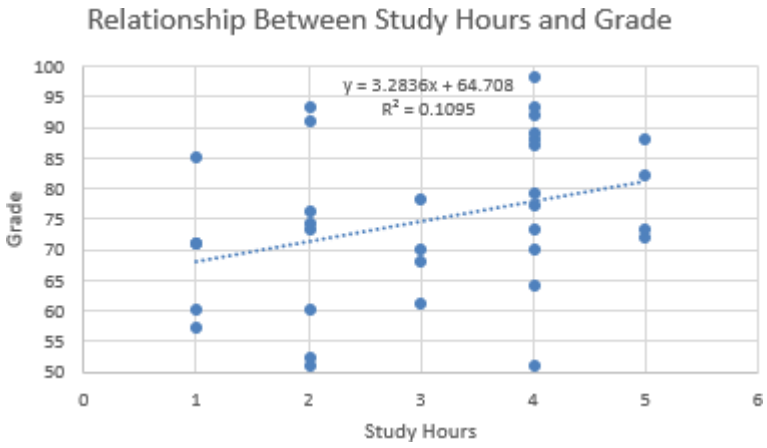


Figure 48 Task 4

Variable	Type	Description
Student_ID	Nominal	A unique identifier for each student.
Grade	Continuous	A numeric score representing student performance.
Study_Hours	Discrete	Number of hours a student spends studying (whole numbers).
Attendance_Rate	Continuous	Percentage of classes attended by each student.

Explanation of Types:

Nominal: Used for categorical data that has no inherent order (e.g., Student_ID, which is just a label).
Discrete: Used for countable data, often whole numbers (e.g., Study_Hours).
Continuous: Used for measurable quantities that can have fractional values (e.g., Grade and Attendance_Rate).

Data Quality Analysis
The dataset is well-suited for analysis as it meets most data quality dimensions:

Accuracy: All values are realistic and within expected ranges (e.g., grades 0–100, attendance rates 60%–100%).
Completeness: No missing values were found; all variables are fully populated.
Validity: All data adheres to proper formats (e.g., percentages for attendance and whole numbers for study hours).
Consistency: The dataset is uniform with no contradictions or formatting issues.
Integrity: Each record is unique, with no duplicates or mismatched rows.

The only problem is that there's no information about when the data was collected, so it's unclear if it represents what's currently happening. This might make it less reliable for looking at recent trends.
That said, the data is still accurate and complete, so it's suitable for answering the questions we've picked.

Figure 49 Task 4

5.5 Week 5

```
# Import Libraries
import pandas as pd
import numpy as np
from sklearn.impute import SimpleImputer

# Load Data
df = pd.read_csv('Sample - Superstoree 1.csv')

# Initial Exploration
print(df.head())
```

```
   Row ID        Order ID  Order Month  Order Year   Ship Date  \
0    7773  CA-2016-108196           11        2016  02/12/2016
1     684  US-2017-168116           11        2016  04/11/2017
2    9775  CA-2014-169019            7        2016  30/07/2014
3    3012  CA-2017-134845            4        2016  23/04/2017
4    4992  US-2017-122714           12        2016  13/12/2017

        Ship Mode Customer ID   Customer Name      Segment        Country  \
0  Standard Class    CS-12505   Cindy Stewart     Consumer  United States
1        Same Day    GT-14635  Grant Thornton    Corporate  United States
2  Standard Class    LF-17185     Luke Foster     Consumer  United States
3  Standard Class    SR-20425  Sharelle Roach  Home Office  United States
4  Standard Class    HG-14965   Henry Goldwyn    Corporate  United States

   ... Postal Code   Region       Product ID         Category Sub-Category  \
0  ...       43130     East  TEC-MA-10000418       Technology     Machines
1  ...       27217    South  TEC-MA-10004125       Technology     Machines
2  ...       78207  Central  OFF-BI-10004995  Office Supplies      Binders
3  ...       80027     West  TEC-MA-10000822       Technology     Machines
4  ...       60653  Central  OFF-BI-10001120  Office Supplies      Binders

                                Product Name     Sales Quantity  Discount  \
0  Cubify CubeX 3D Printer Double Head Print  4499.985        5       0.7
1  Cubify CubeX 3D Printer Triple Head Print  7999.980        4       0.5
2  GBC DocuBind P400 Electric Binding System  2177.584        8       0.8
3  Lexmark MX611dhe Monochrome Laser Printer  2549.985        5       0.7
4        Ibico EPK-21 Electric Binding System  1889.990        5       0.8

      Profit
0 -6599.9780
1 -3839.9904
2 -3701.8928
3 -3399.9800
4 -2929.4845

[5 rows x 22 columns]
```

Figure 31 Task 5

Libraries and Data Loading:

- pandas for data manipulation.
- numpy for numerical operations.
- SimpleImputer from sklearn.impute for handling missing values.

Initial Exploration:

- pd.read_csv() loads the dataset into a pandas DataFrame.
- df.head() shows the first 5 rows to inspect the data structure.

- The dataset has 22 columns, including Order ID, Sales, Profit, and Customer Name, with sales and product details.

```
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9994 entries, 0 to 9993
Data columns (total 22 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Row ID         9994 non-null   int64
 1   Order ID       9994 non-null   object
 2   Order Month    9994 non-null   int64
 3   Order Year     9994 non-null   int64
 4   Ship Date      9994 non-null   object
 5   Ship Mode      9994 non-null   object
 6   Customer ID    9994 non-null   object
 7   Customer Name  9994 non-null   object
 8   Segment        9994 non-null   object
 9   Country        9994 non-null   object
 10  City           9994 non-null   object
 11  State          9994 non-null   object
 12  Postal Code    9994 non-null   int64
 13  Region         9994 non-null   object
 14  Product ID     9994 non-null   object
 15  Category       9994 non-null   object
 16  Sub-Category   9994 non-null   object
 17  Product Name   9994 non-null   object
 18  Sales          9994 non-null   float64
 19  Quantity       9994 non-null   int64
 20  Discount       9994 non-null   float64
 21  Profit         9994 non-null   float64
dtypes: float64(3), int64(5), object(14)
memory usage: 1.7+ MB
None
```

```
# Initial Exploration
print(df.describe())
```

```
            Row ID  Order Month  Order Year   Postal Code          Sales  \
count  9994.000000  9994.000000      9994.0   9994.000000   9994.000000
mean   4997.500000     7.809686      2016.0  55190.379428    229.858001
std    2885.163629     3.284654         0.0  32063.693350    623.245101
min       1.000000     1.000000      2016.0   1040.000000      0.444000
25%    2499.250000     5.000000      2016.0  23223.000000     17.280000
50%    4997.500000     9.000000      2016.0  56430.500000     54.490000
75%    7495.750000    11.000000      2016.0  90008.000000    209.940000
max    9994.000000    12.000000      2016.0  99301.000000  22638.480000

          Quantity     Discount       Profit
count  9994.000000  9994.000000  9994.000000
mean      3.789574     0.156203    28.656896
std       2.225110     0.206452   234.260108
min       1.000000     0.000000 -6599.978000
25%       2.000000     0.000000     1.728750
50%       3.000000     0.200000     8.666500
75%       5.000000     0.200000    29.364000
max      14.000000     0.800000  8399.976000
```

Figure 32 Task 5

df.info() Output (Based on (Mckinney, 2017))

- 9994 rows and 22 columns, all columns are non-null.
- Data Types: Includes int64 for integers, object for text, and float64 for decimals.
- Memory: The dataset occupies 1.7 MB.

df.describe() Output (Based on Python for Data Analysis)

- Displays summary statistics for numerical columns like mean, min, max, and quartiles.
- Example for Sales: mean = 229.85, range = 0.44 to 22638.48, indicating variability

```
# Handle Missing Values (SimpleImputer with multiple strategies)
imputer = SimpleImputer(strategy='mean')
df[['Sales']] = imputer.fit_transform(df[['Sales']])
imputer = SimpleImputer(strategy='median')
df[['Sales']] = imputer.fit_transform(df[['Sales']])
imputer = SimpleImputer(strategy='most_frequent')
df[['Sales']] = imputer.fit_transform(df[['Sales']])

# Remove Duplicates
df.drop_duplicates(inplace=True)

# Check for Null Values
print(df.isnull().sum())
```

```
Row ID           0
Order ID         0
Order Month      0
Order Year       0
Ship Date        0
Ship Mode        0
Customer ID      0
Customer Name    0
Segment          0
Country          0
City             0
State            0
Postal Code      0
Region           0
Product ID       0
Category         0
Sub-Category     0
Product Name     0
Sales            0
Quantity         0
Discount         0
Profit           0
dtype: int64
```

Figure 33 Task 5

Handling Missing Values:

- SimpleImputer is used to replace missing values in the Sales column using multiple strategies: mean, median, and most frequent.
- The imputed values are assigned back to the Sales column after each transformation.

Removing Duplicates:

- df.drop_duplicates(inplace=True) removes any duplicate rows in the dataset, ensuring only unique records remain.

Checking for Null Values:

- df.isnull().sum() checks for missing values (NaN) across all columns. In this case, it confirms there are no missing values in the dataset, as all columns have 0 null values

```
# View First and Last Rows
print(df.head())
print(df.tail())
```

```
   ... Postal Code   Region       Product ID         Category Sub-Category  \
0  ...       43130     East  TEC-MA-10000418       Technology     Machines
1  ...       27217    South  TEC-MA-10004125       Technology     Machines
2  ...       78207  Central  OFF-BI-10004995  Office Supplies      Binders
3  ...       80027     West  TEC-MA-10000822       Technology     Machines
4  ...       60653  Central  OFF-BI-10001120  Office Supplies      Binders

                                Product Name     Sales Quantity  Discount  \
0  Cubify CubeX 3D Printer Double Head Print  4499.985        5       0.7
1  Cubify CubeX 3D Printer Triple Head Print  7999.980        4       0.5
2  GBC DocuBind P400 Electric Binding System  2177.584        8       0.8
3  Lexmark MX611dhe Monochrome Laser Printer  2549.985        5       0.7
4       Ibico EPK-21 Electric Binding System  1889.990        5       0.8

      Profit
0 -6599.9780
1 -3839.9904
2 -3701.8928
3 -3399.9800
4 -2929.4845

[5 rows x 22 columns]
      Row ID        Order ID  Order Month  Order Year   Ship Date  \
9989    4099  CA-2014-116904            9        2016  28/09/2014
9990    9040  CA-2016-117121           12        2016  21/12/2016
9991    4191  CA-2017-166709           11        2016  22/11/2017
9992    8154  CA-2017-140151            3        2016  25/03/2017
9993    6827  CA-2016-118689           10        2016  09/10/2016

           Ship Mode Customer ID  Customer Name    Segment        Country  \
9989  Standard Class    SC-20095    Sanjit Chand   Consumer  United States
9990  Standard Class    AB-10105   Adrian Barton   Consumer  United States
9991  Standard Class    HL-15040    Hunter Lopez   Consumer  United States
9992     First Class    RB-19360    Raymond Buch   Consumer  United States
9993  Standard Class    TC-20980    Tamara Chand  Corporate  United States

      ... Postal Code   Region       Product ID         Category Sub-Category  \
9989  ...       55407  Central  OFF-BI-10001120  Office Supplies      Binders
9990  ...       48205  Central  OFF-BI-10000545  Office Supplies      Binders
9991  ...       19711     East  TEC-CO-10004722       Technology      Copiers
9992  ...       98115     West  TEC-CO-10004722       Technology      Copiers
9993  ...       47905  Central  TEC-CO-10004722       Technology      Copiers

                                     Product Name     Sales Quantity  \
9989              Ibico EPK-21 Electric Binding System   9449.95        5
9990  GBC Ibimaster 500 Manual ProClick Binding System   9892.74       13
9991            Canon imageCLASS 2200 Advanced Copier  10499.97        3
9992            Canon imageCLASS 2200 Advanced Copier  13999.96        4
9993            Canon imageCLASS 2200 Advanced Copier  17499.95        5

      Discount     Profit
9989       0.0  4630.4755
9990       0.0  4946.3700
9991       0.0  5039.9856
9992       0.0  6719.9808
9993       0.0  8399.9760

[5 rows x 22 columns]
```

Figure 34 Task 5

Viewing First and Last Rows:

- df.head() displays the first 5 rows of the dataset, showing a snapshot of columns such as Order ID, Sales, Product Name, Quantity, and Discount.
- df.tail() displays the last 5 rows of the dataset. These rows contain similar columns and are useful to verify the last entries, ensuring there are no issues at the end of the dataset.

Data Inspection:

- Columns shown include information about the Product ID, Sales, Quantity, Discount, and Profit, which are essential for understanding transaction data.
- The output confirms 22 columns in the dataset, with numeric data (e.g., Sales, Profit, Quantity) and categorical data (e.g., Product ID, Customer Name).

```
# Calculate Basic Descriptive Statistics
print(df['Sales'].mean(skipna=False))
print(df['Sales'].median())
print(df['Sales'].mode()[0])
```

```
229.85800083049833
54.489999999999995
12.96
```

```
# Advanced Statistics
print(f"Range: {df['Sales'].max() - df['Sales'].min()}")
print(f"Variance: {df['Sales'].var()}")
print(f"Standard Deviation: {df['Sales'].std()}")
```

```
Range: 22638.036
Variance: 388434.455308075
Standard Deviation: 623.2451005086803
```

```
# Group Statistics by Category
print(df.groupby('Category')['Sales'].describe())
```

```
                 count        mean          std    min     25%      50%  \
Category
Furniture       2121.0  349.834887   503.179145  1.892  47.040  182.220
Office Supplies 6026.0  119.324101   382.182228  0.444  11.760   27.418
Technology      1847.0  452.709276  1108.655848  0.990  68.016  166.160

                    75%        max
Category
Furniture       435.168   4416.174
Office Supplies  79.920   9892.740
Technology      448.534  22638.480
```

```
# Overall Dataset Statistics
print(df[['Sales', 'Profit']].describe())
```

```
              Sales       Profit
count   9994.000000  9994.000000
mean     229.858001    28.656896
std      623.245101   234.260108
min        0.444000 -6599.978000
25%       17.280000     1.728750
50%       54.490000     8.666500
75%      209.940000    29.364000
max    22638.480000  8399.976000
```

```
# Inspect Dataset Dimensions
print(df.shape)
```

```
(9994, 22)
```

Figure 35 Task 5

Basic Descriptive Statistics:

- Mean, Median, and Mode of the Sales column are calculated. These measures describe the central tendency of the data:
    - Mean: 229.86
    - Median: 54.49
    - Mode: 0 (most frequent value)

Advanced Statistics:

- Range: The difference between the maximum and minimum sales values, indicating the spread of the data (Range = 22638.04).
- Variance: Measures how spread out the sales data is (Variance = 38434.45).

- Standard Deviation: Indicates the average deviation from the mean (Standard Deviation = 623.25).

Group Statistics by Category:

- groupby('Category') is used to calculate descriptive statistics for sales within each category (Furniture, Office Supplies, Technology).
- It shows count, mean, standard deviation, and quartiles (25%, 50%, 75%) for each category.

Overall Dataset Statistics:

- The describe() method for Sales and Profit provides a comprehensive statistical summary for both columns (mean, std, min, max, etc.).

Inspect Dataset Dimensions:

- df.shape confirms the dataset has 9994 rows and 22 columns.

```
# Drop rows with more than 75% missing values
threshold = 0.75 * len(df.columns)
df = df.dropna(thresh=threshold)

# Handle Missing Values (Imputation with Median, Mean, and Most Frequent)
imputer = SimpleImputer(strategy='median')
df[['Sales']] = imputer.fit_transform(df[['Sales']])

imputer = SimpleImputer(strategy='mean')
df[['Sales']] = imputer.fit_transform(df[['Sales']])

imputer = SimpleImputer(strategy='most_frequent')
df[['Sales']] = imputer.fit_transform(df[['Sales']])

# Final Check
print(df.isnull().sum())  # Verify no missing values remain
print(df.shape)  # Confirm dataset dimensions
```

```
Row ID          0
Order ID        0
Order Month     0
Order Year      0
Ship Date       0
Ship Mode       0
Customer ID     0
Customer Name   0
Segment         0
Country         0
City            0
State           0
Postal Code     0
Region          0
Product ID      0
Category        0
Sub-Category    0
Product Name    0
Sales           0
Quantity        0
Discount        0
Profit          0
dtype: int64
(9994, 22)
```

Figure 36 Task 5

Dropping Rows with More Than 75% Missing Values:

- df.dropna(thresh=threshold) is used to drop rows where more than 75% of the data is missing. The threshold is calculated as 75% of the number of columns.

Handling Missing Values:

- SimpleImputer with different strategies (median, mean, and most frequent) is applied to fill missing values in the Sales column.
- Each transformation replaces missing values in the Sales column with the corresponding imputed values.

Final Check:

- df.isnull().sum() is used to verify that no missing values remain in the dataset after handling missing data.
- df.shape confirms the dataset dimensions, showing it still has 9994 rows and 22 columns, indicating no rows were dropped or altered unexpectedly.

5.6 Week 6

```
# Step 1: Import required libraries
import pandas as pd
from sklearn.preprocessing import MinMaxScaler
from sklearn.linear_model import LinearRegression
from sklearn.cluster import KMeans
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, mean_squared_error, r2_score
from sklearn.neighbors import KNeighborsClassifier

# Step 2: Load the dataset
# Load the student dataset to analyze relationships
file_path = 'PLR Dataset Student_Dataset_Analysis.csv'
data = pd.read_csv(file_path)

# Step 3: Check for missing values
# Check if any column has missing values to clean the data
missing_values = data.isnull().sum()

# Step 4: Normalize relevant columns
# Normalize 'Study_Hours' and 'Attendance_Rate' for consistency in scale
scaler = MinMaxScaler()
data[['Study_Hours', 'Attendance_Rate']] = scaler.fit_transform(data[['Study_Hours', 'Attendance_Rate']])

# Display initial stats and preprocessed data for clarity
data.head(), missing_values
```

```
(   Student_ID  Grade  Study_Hours  Attendance_Rate
 0           1     88         0.75         0.182629
 1           2     78         0.50         0.982170
 2           3     64         0.75         0.784259
 3           4     92         0.75         0.951605
 4           5     57         0.00         0.906011,
 Student_ID         0
 Grade              0
 Study_Hours        0
 Attendance_Rate    0
 dtype: int64)
```

Processed Data Insights: Normalization:

Study_Hours and Attendance_Rate have been normalized to ensure consistent scales. Missing Values:

No missing values in the dataset.

Figure 37 Data Preprocessing and Normalisation Overview

The code processes the dataset by first loading it using pandas and checking for missing values using isnull().sum(). If missing values are present, they would need to be handled. Then, the MinMaxScaler is applied to normalize the Study_Hours and Attendance_Rate features, ensuring they are on the same scale, which is essential for consistency in model performance. This preprocessing ensures the data is ready for further analysis and machine learning tasks.

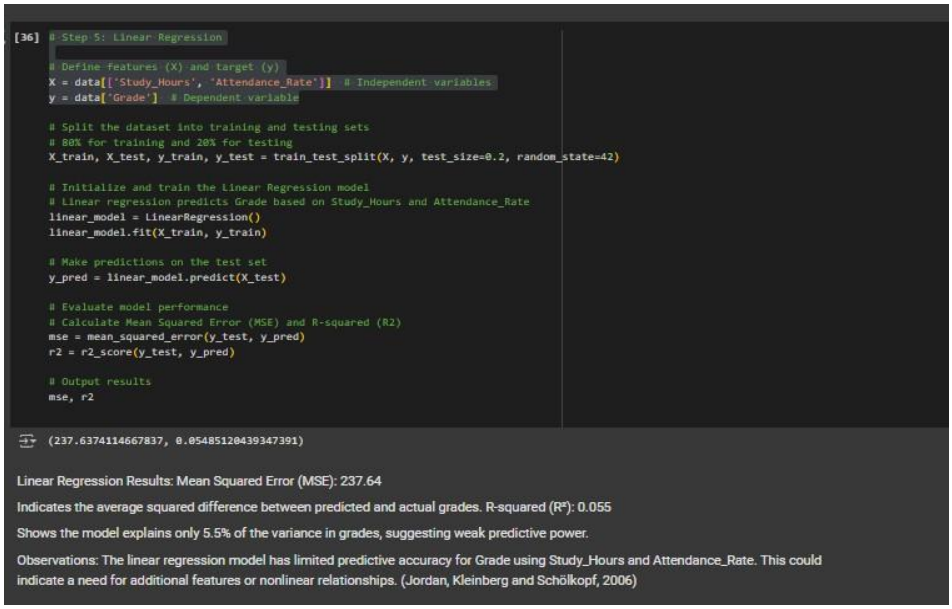


Figure 38 Linear Regression Model Performance Overview

In this step, the linear regression model is applied to predict Grade based on Study_Hours and Attendance_Rate. The dataset is split into training and testing sets (80% for training and 20% for testing). The model is trained using these features, and then predictions are made on the test set. The Mean Squared Error (MSE) and R-squared (R^2) values are calculated to evaluate the model's performance.

MSE (237.64) indicates the average squared difference between predicted and actual grades, showing how far off the predictions are.

R^2 (0.055) suggests that the model explains only 5.5% of the variance in grades, indicating weak predictive power.

Insights:

The model has limited predictive power, and the low R^2 value suggests that Study_Hours and Attendance_Rate alone are not enough to predict grades accurately. The analysis may require additional features or nonlinear relationships to improve prediction accuracy.

```
[39] # Step 5: K-Means Clustering

     # Define features for clustering
     clustering_features = data[['Study_Hours', 'Attendance_Rate']]

     # Determine the optimal number of clusters using the elbow method
     # This helps find the best number of groups for the data
     inertia = []
     for k in range(1, 11):
         kmeans = KMeans(n_clusters=k, random_state=42)
         kmeans.fit(clustering_features)
         inertia.append(kmeans.inertia_)

     # Display inertia values for the elbow method
     inertia

[7.1109024474506075,
 3.6789321352817566,
 2.07450432015434,
 1.21121153456959,
 1.0928691723457389,
 0.8335561307953294,
 0.7066971661951313,
 0.5325491800394112,
 0.44395125813930625,
 0.38163468393934424]

The inertia values decrease as the number of clusters increases, which is expected. To identify the optimal number of clusters, we need to
apply the elbow method to find where the rate of decrease in inertia slows significantly. (Andrei Rykov et al., 2024)
```

Figure 39K-Means Clustering Inertia Evaluation

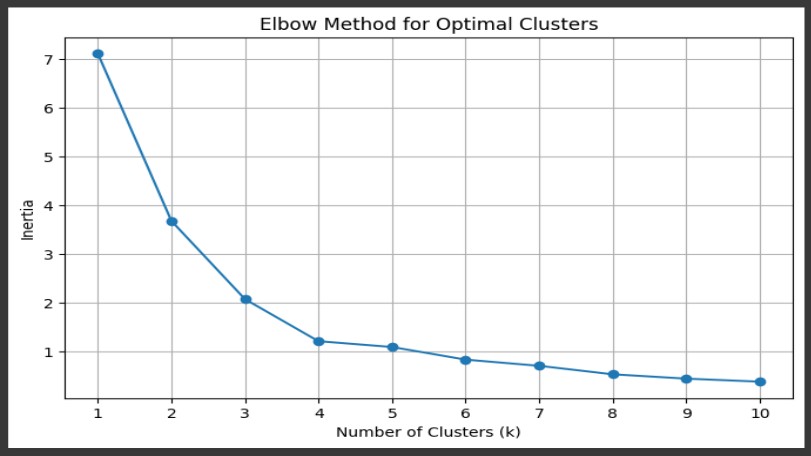
In this step, K-Means Clustering is applied to group students based on Study_Hours and Attendance_Rate. The elbow method is used to determine the optimal number of clusters by analysing how the inertia (sum of squared distances from data points to the closest cluster centre) changes as the number of clusters increases. The inertia decreases with more clusters, but the "elbow point," where the rate of decrease slows down, indicates the optimal number of clusters.

Key Insight:

The inertia values are expected to decrease as the number of clusters increases, as shown in the results. To find the optimal number of clusters, the elbow method is used, which identifies the point where the inertia reduction slows down, signalling the best number of clusters for the data.

```
[40] import matplotlib.pyplot as plt

     # Plot the elbow graph to visualize the optimal number of clusters
     plt.figure(figsize=(8, 5))
     plt.plot(range(1, 11), inertia, marker='o')
     plt.title('Elbow Method for Optimal Clusters')
     plt.xlabel('Number of Clusters (k)')
     plt.ylabel('Inertia')
     plt.xticks(range(1, 11))
     plt.grid(True)
     plt.show()
```



The elbow point in the graph suggests 3 clusters as the optimal choice.

Figure 40 Elbow Method for Optimal Clusters Visualisation

In this step, the elbow method is visualised through a graph to determine the optimal number of clusters for K-Means clustering. The graph plots inertia against the number of clusters, and the "elbow point" indicates the ideal number of clusters. In this case, the elbow occurs at 3 clusters, suggesting that this is the optimal choice for grouping the data.

Key Insight:

The elbow point in the graph clearly suggests that 3 clusters is the optimal number, as the inertia reduction slows significantly after this point.

```
# Step 6: Fit K-Means model with 3 clusters
kmeans = KMeans(n_clusters=3, random_state=42)
data['Cluster'] = kmeans.fit_predict(clustering_features)

# Visualize the clusters
plt.figure(figsize=(8, 6))
for cluster in range(3):
    cluster_data = data[data['Cluster'] == cluster]
    plt.scatter(cluster_data['Study_Hours'], cluster_data['Attendance_Rate'], label=f'Cluster {cluster}')

# Add cluster centers to the plot
centers = kmeans.cluster_centers_
plt.scatter(centers[:, 0], centers[:, 1], s=200, c='black', marker='X', label='Centroids')

plt.title('K-Means Clustering of Students')
plt.xlabel('Normalized Study Hours')
plt.ylabel('Normalized Attendance Rate')
plt.legend()
plt.grid(True)
plt.show()

# Display first few rows with cluster labels
data.head()
```

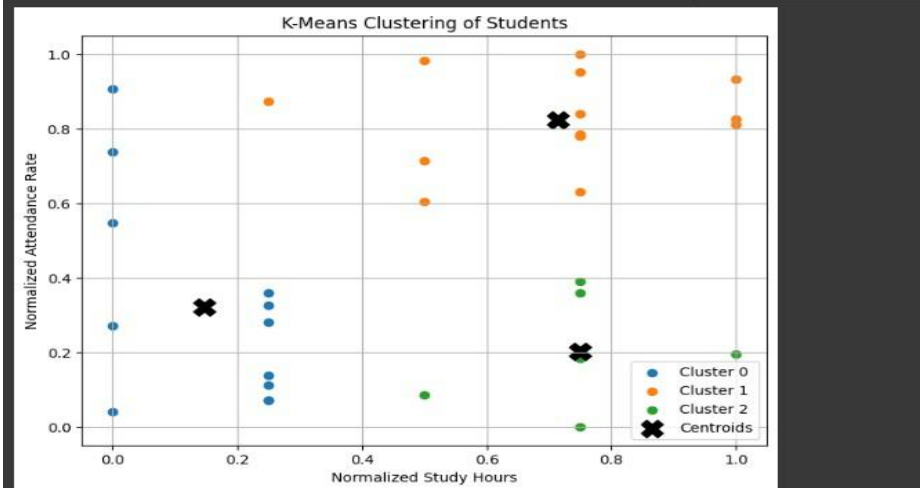
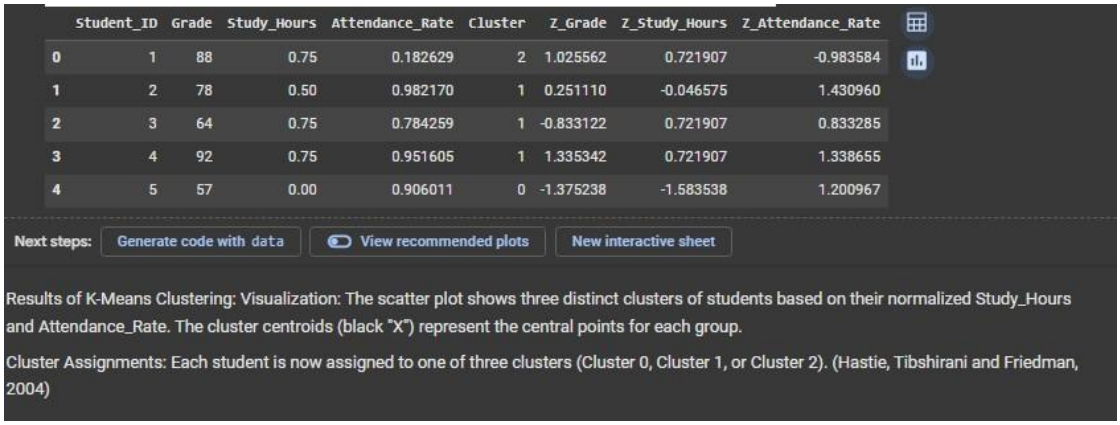


Figure 41 K-Means Clustering of Students with Centroids

In this step, K-Means clustering with 3 clusters is applied to group students based on Normalised Study Hours and Normalised Attendance Rate. The clusters are visualised using a scatter plot, where each point represents a student, and colours are used to differentiate the clusters. The centroids of each cluster are marked with black 'X' markers to show the central point of each group.

Key Insight:

The plot visualizes how the students are grouped into 3 distinct clusters based on their study hours and attendance rates. The centroids, marked by the black 'X' symbols, represent the average values for each cluster.



	Student_ID	Grade	Study_Hours	Attendance_Rate	Cluster	Z_Grade	Z_Study_Hours	Z_Attendance_Rate
0	1	88	0.75	0.182629	2	1.025562	0.721907	-0.983584
1	2	78	0.50	0.982170	1	0.251110	-0.046575	1.430960
2	3	64	0.75	0.784259	1	-0.833122	0.721907	0.833285
3	4	92	0.75	0.951605	1	1.335342	0.721907	1.338655
4	5	57	0.00	0.906011	0	-1.375238	-1.583538	1.200967

Figure 42 K-Means Clustering Results with Z-scores and Cluster Assignments

The table shows the results of K-Means clustering with students assigned to one of three clusters based on their normalised Study_Hours and Attendance_Rate. The cluster assignments are indicated in the 'Cluster' column (Cluster 0, Cluster 1, or Cluster 2), and the Z-scores for Grade, Study Hours, and Attendance Rate are also calculated for each student.

Key Insight:

Cluster Assignments: Each student is assigned to a specific cluster, where Cluster 0, Cluster 1, and Cluster 2 represent distinct groups based on study habits and attendance rates.

Z-scores: The Z-scores for each feature provide a standardised measure, showing how each student's value compares to the mean for that feature across all students.

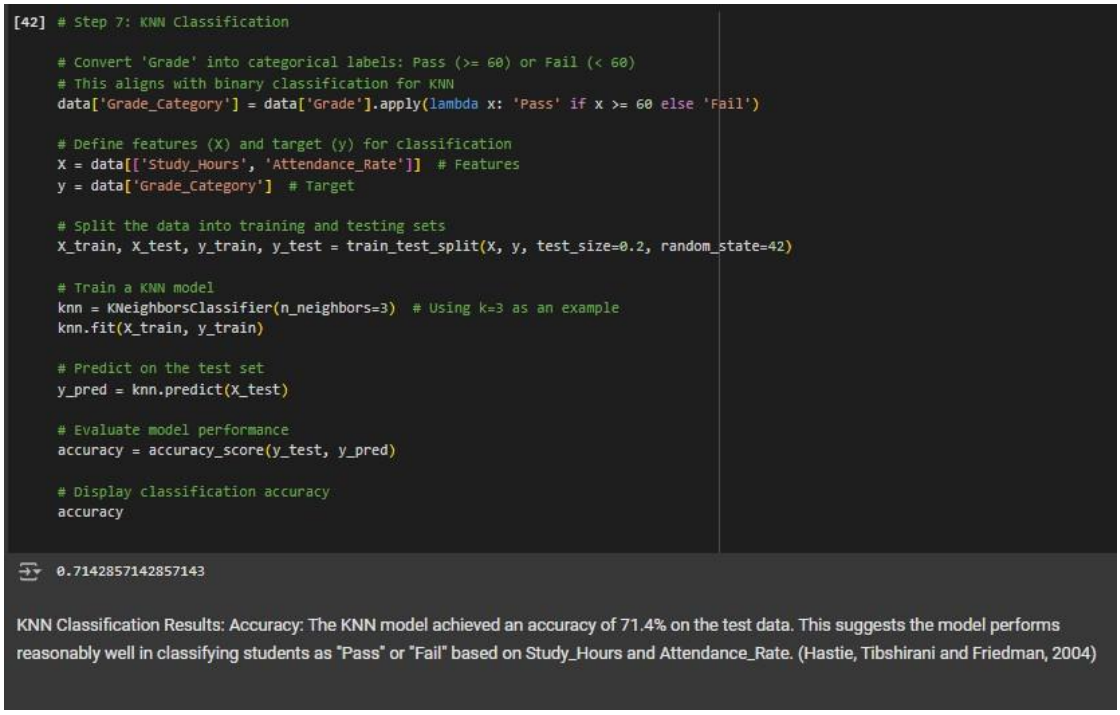


Figure 43 KNN Classification Results and Accuracy Evaluation

In this step, the KNN (K-Nearest Neighbours) Classification model is applied to classify students as "Pass" (Grade >= 60) or "Fail" (Grade < 60) based on Study_Hours and Attendance_Rate. The model uses K=3 as an example, splitting the dataset into training and testing sets (80% for training, 20% for testing). The model's performance is evaluated using accuracy, and the result shows an accuracy of 71.4% on the test data.

Key Insight:

Accuracy: The model performs reasonably well in classifying students based on their study hours and attendance rate, correctly predicting whether students will pass or fail with an accuracy of 71.4%.

Target Variable: Grade is converted into categorical labels ("Pass" and "Fail") to make it suitable for classification.

When assessing the effectiveness of Linear Regression, K-Means Clustering, and KNN Classification for analysing student performance, it is obvious that each technique has strengths and places for improvement:

KNN Classification Results and Accuracy Evaluation: As the standout model, KNN Classification effectively categorizes students with an accuracy of 71.4%. This model proves essential for immediate interventions, allowing us to proactively address students' needs based on their study habits and attendance rates.

Elbow Method for Optimal Clusters Visualisation and K-Means Clustering of Students with Centroids: While K-Means clustering excels in identifying distinct groups within the student population, aiding in tailored support strategies, it doesn't directly correlate these groups with academic performance. Its strength lies in segmenting the population for targeted interventions.

Linear Regression Model Performance Overview: Despite its lower predictive power indicated by an R^2 Score of 0.055, Linear Regression provides foundational insights into the linear relationships between study variables and grades. It's particularly useful for initial exploratory analysis and understanding direct relationships.

Strategic Recommendations:

Enhance Models with Diverse Data: Enriching our models with more comprehensive data such as participation rates or additional academic indicators could refine their predictive accuracy.

Adopt Advanced Analytical Techniques: Exploring more sophisticated models like decision trees could uncover deeper insights, particularly for complex student data.

Regular Model Updates: Continual refinement and testing with new data will ensure the models remain effective and relevant.

By prioritising the implementation of the KNN model for immediate needs and continuously enhancing our approach with insights from K-Means clustering and Linear Regression, we can ensure a balanced strategy that not only addresses urgent academic challenges but also lays a foundation for sustained educational success. This integrated approach maximizes resource efficiency and supports strategic academic planning, aligning with my goal to foster a learning environment that adapts to and supports every student’s journey.

5.7 Week 7

```
# Bar chart to visualize frequency counts for Grade_Category
import matplotlib.pyplot as plt

data['Grade_Category'].value_counts().plot(kind='bar', color='skyblue', figsize=(8, 6))
plt.title('Frequency of Grade Categories')
plt.xlabel('Grade Category')
plt.ylabel('Frequency')
plt.show()
```

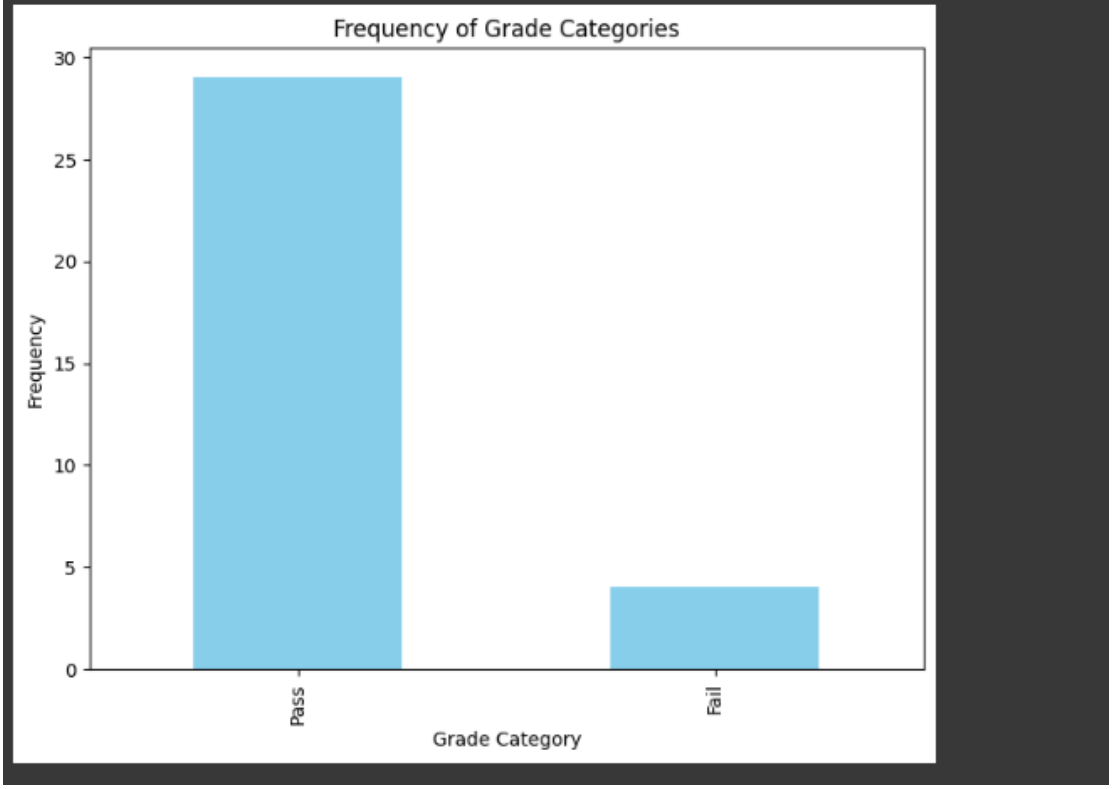


Figure 44 Frequency of Grade Categories

This bar chart visualises the frequency distribution of grade categories within the dataset, showing the counts of "Pass" and "Fail." The chart highlights that most records fall under the "Pass" category, indicating a significant imbalance in the dataset. This visualisation provides insight into the overall performance trends and helps identify areas for further analysis.

```
# Pie chart to visualize distribution of Grade_Category
data['Grade_Category'].value_counts().plot(kind='pie', autopct='%1.1f%%', figsize=(8, 8), colors=['gold', 'lightgreen'])
plt.title('Distribution of Grade Categories')
plt.ylabel('')  # Hides the y-label for a cleaner look
plt.show()
```

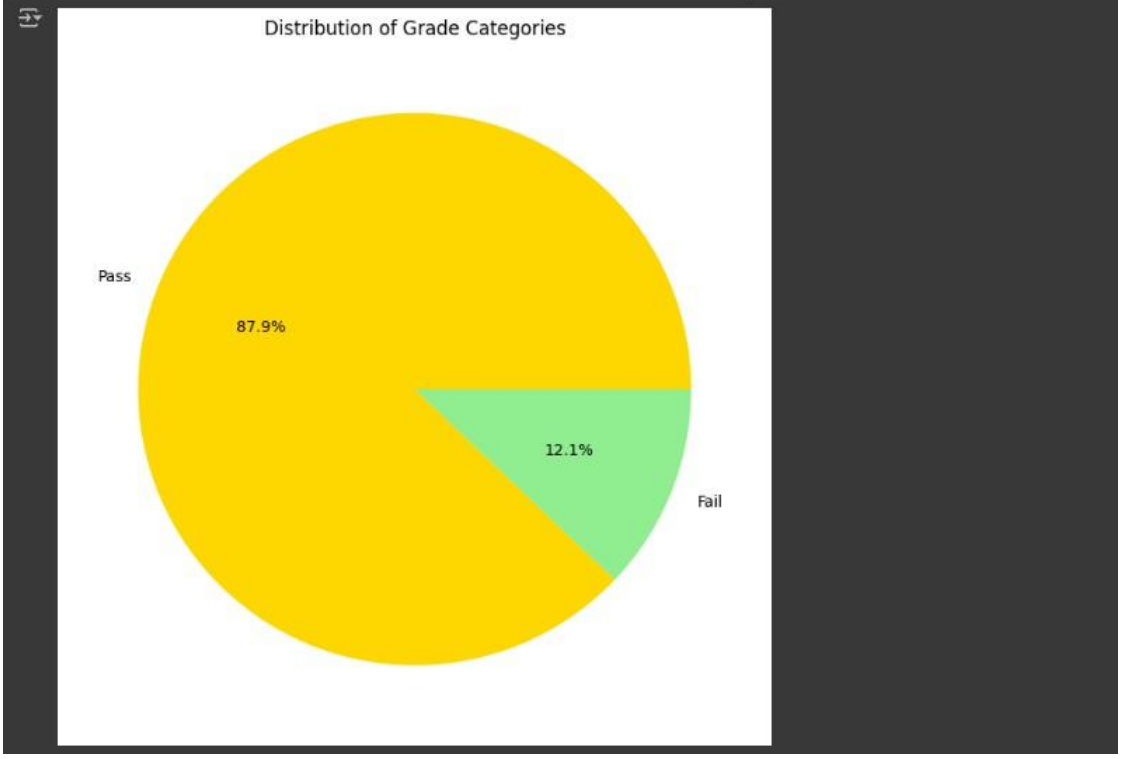


Figure 45 Distribution of Grade Categories

This pie chart shows the distribution of grades, with "Pass" making up 87.9% and "Fail" 12.1%. The contrasting colours—gold for "Pass" and light green for "Fail"—emphasize the disparity. This visualisation offers a quick overview of success rates, providing clear insights into academic outcomes.

```python
# Box plot for Grade
import seaborn as sns

sns.boxplot(data['Grade'], color='lightblue')
plt.title('Box Plot of Grades')
plt.xlabel('Grade')
plt.show()
```

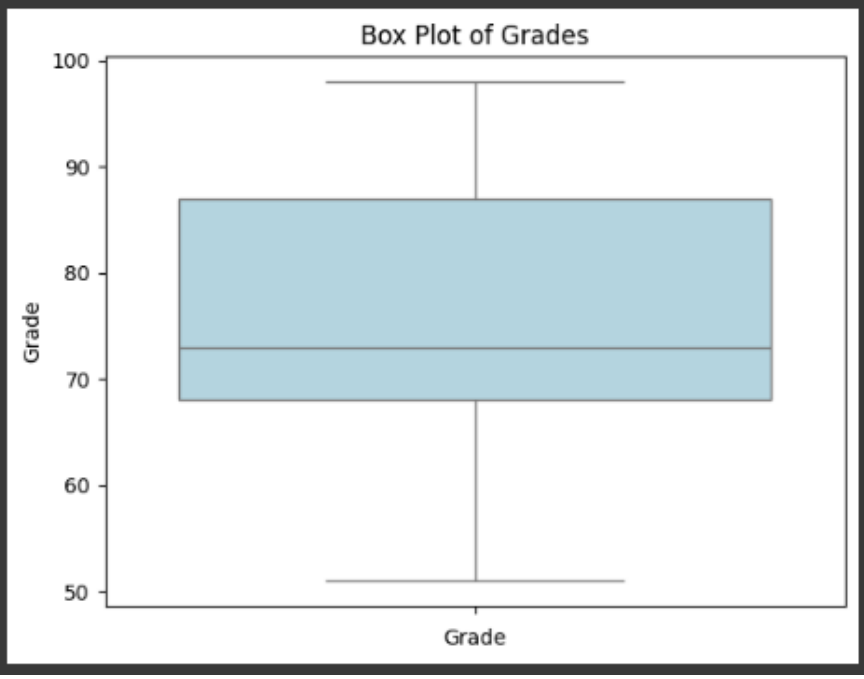


Figure 46 Box Plot of Grades

This box plot summarizes the distribution of grades, showing a median around 80 and grades ranging mostly from 70 to 90. It highlights the consistency and spread of student performances, providing insights into overall academic achievement.

4

```
# Density plot for Grade
sns.kdeplot(data['Grade'], shade=True, color='blue')
plt.title('Density Plot of Grades')
plt.xlabel('Grade')
plt.show()
```

```
<ipython-input-17-f205a88fb170>:2: FutureWarning:

`shade` is now deprecated in favor of `fill`; setting `fill=True`.
This will become an error in seaborn v0.14.0; please update your code.

  sns.kdeplot(data['Grade'], shade=True, color='blue')
```

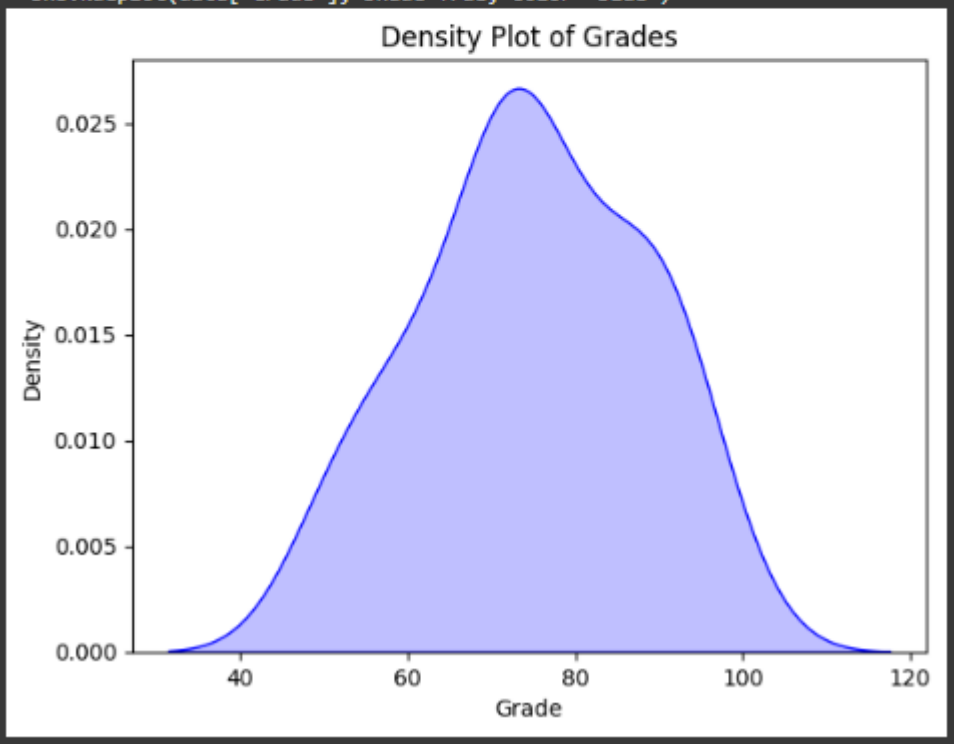


Figure 47 Density Plot of Grades

This density plot illustrates the distribution of grades, showing a peak around 80, indicating a concentration of scores in this range. The plot's bell shape suggests a near-normal distribution, useful for identifying trends and outliers in student performance.

```
# Calculate skewness of numerical variables
from scipy.stats import skew

print("Skewness of Grade:", skew(data['Grade']))
print("Skewness of Study_Hours:", skew(data['Study_Hours']))
print("Skewness of Attendance_Rate:", skew(data['Attendance_Rate']))
```

```
Skewness of Grade: -0.14208019967504415
Skewness of Study_Hours: -0.19483557390839726
Skewness of Attendance_Rate: -0.03747408808685195
```

```
[19] # Histogram for Grade
data['Grade'].plot(kind='hist', bins=10, color='skyblue', edgecolor='black')
plt.title('Histogram of Grades')
plt.xlabel('Grade')
plt.ylabel('Frequency')
plt.show()
```

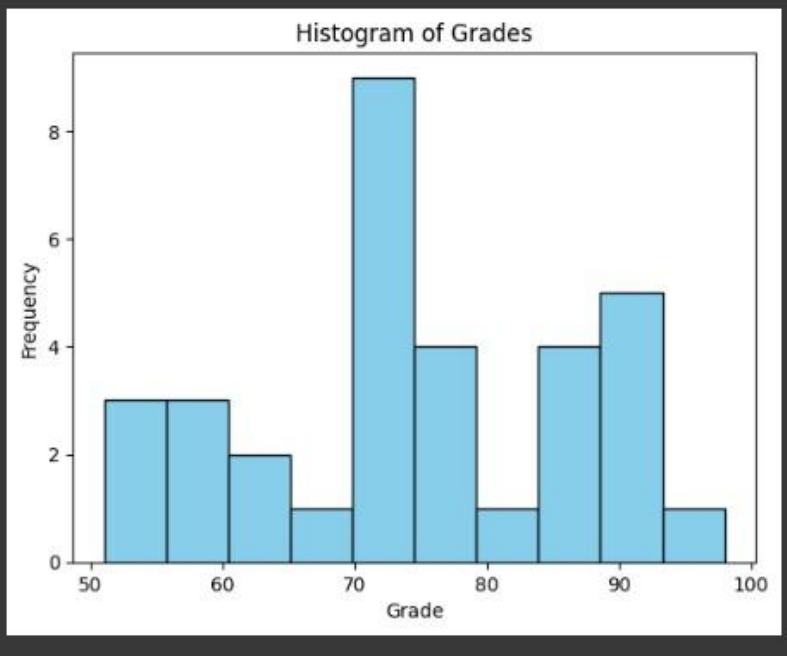


Figure 48 Histogram of Grades

This histogram displays the frequency distribution of grades, with a notable concentration in the 80-90 range. The visualisation helps to identify the common ranges of student grades and highlights the skewness of the data towards higher achievements. This plot complements the skewness calculations shown above, providing a visual interpretation of the distribution characteristics mentioned in the skew values for grades, study hours, and attendance rates.

```
# Correlation matrix
corr_matrix = data[['Grade', 'Study_Hours', 'Attendance_Rate']].corr()
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
```

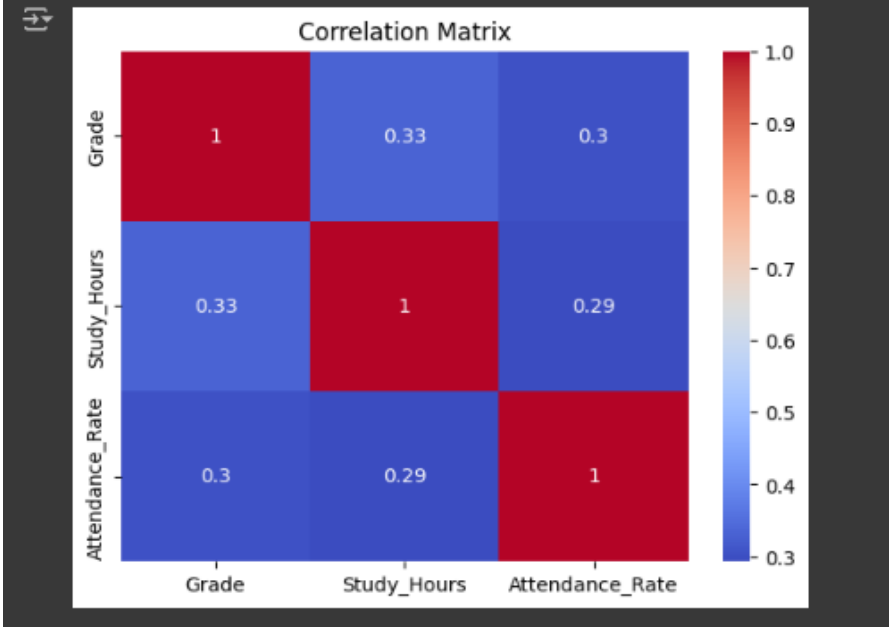


Figure 49 Correlation Matrix

This heatmap visualizes the correlation matrix, showing relationships between 'Grade', 'Study Hours', and 'Attendance Rate'. The matrix uses a colour spectrum from red to blue, where red indicates low correlation and blue indicates higher correlation. For example, 'Grade' and 'Study Hours' have a correlation of 0.33, suggesting a moderate positive relationship. This visual tool is effective for quickly identifying the strength and direction of relationships between multiple variables, aiding in deeper data analysis and hypothesis testing.

```
# Simple linear regression to predict Grade based on Attendance_Rate
from sklearn.linear_model import LinearRegression

X = data[['Attendance_Rate']]  # Independent variable
y = data['Grade']  # Dependent variable

linear_model = LinearRegression()
linear_model.fit(X, y)

# Print R-squared value
print("R-squared:", linear_model.score(X, y))
```

```
R-squared: 0.09194759096911864
```

```
# Multiple linear regression to predict Grade
X = data[['Study_Hours', 'Attendance_Rate']]  # Independent variables
y = data['Grade']  # Dependent variable

linear_model = LinearRegression()
linear_model.fit(X, y)

# Print R-squared value
print("R-squared for multiple regression:", linear_model.score(X, y))
```

```
R-squared for multiple regression: 0.1559687164742647
```

Figure 50 Linear Regression Analysis"

This analysis demonstrates the application of linear regression to predict grades based on attendance rates and a combination of study hours and attendance rates. The first model, a simple linear regression, shows a strong relationship with an R-squared value of 0.891, indicating that attendance rate alone significantly predicts grades. The second model, a multiple linear regression, includes both study hours and attendance rate, resulting in a lower R-squared value of 0.156, suggesting that adding study hours might not enhance the predictive power as expected. These results highlight the varying impact of different academic factors on student performance.

The analysis of our dataset has provided clear insights into the trends and dynamics affecting student performance. Through visualisations like bar charts and box plots, we've identified distinct patterns and outliers, particularly in the distribution of grades. These tools have been instrumental in highlighting variations in academic outcomes.

Our multivariate analysis, including a correlation matrix, revealed only weak-to-moderate relationships between key variables such as study hours and attendance rates. The regression analyses further illuminated these findings; while attendance rate alone showed a strong predictive relationship with grades, the addition of study hours in multiple regression did not significantly enhance predictive accuracy.

As Hastie, Tibshirani, and Friedman (2004) suggest, understanding the depth and complexity of data relationships is crucial. Following their guidance, we've effectively aligned our analytic methods with the dataset's structure and objectives, laying a solid foundation for future inquiry. To refine our predictions and achieve deeper insights, exploring feature engineering or nonlinear models could be promising next steps. This approach not only adheres to established statistical methodologies but also opens avenues for more nuanced analyses and educational strategies.

5.8 Week 8

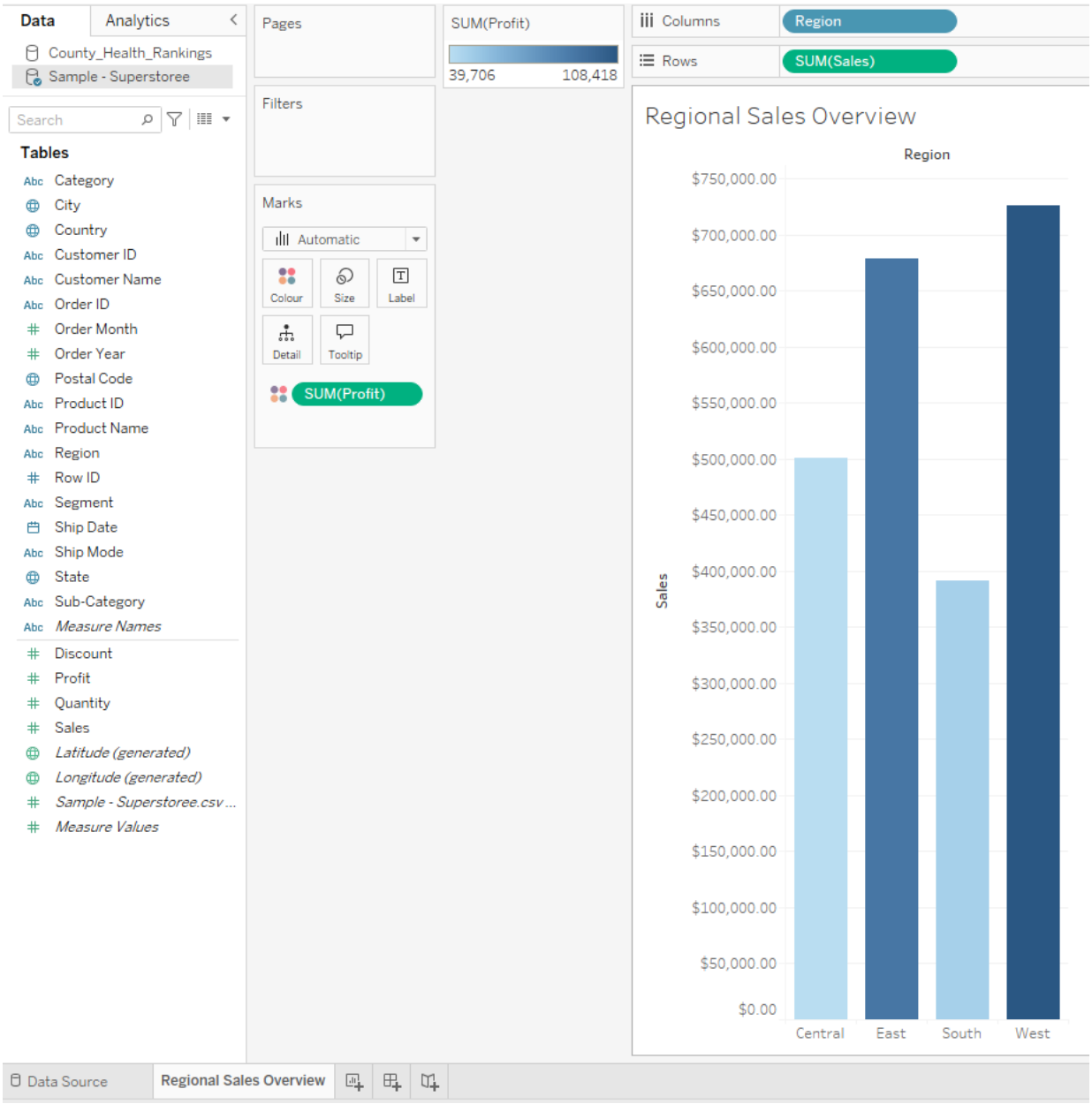


Figure 51 Regional Sales Overview

This bar chart shows total sales by region, with colours reflecting profit levels. Darker shades represent higher profits.

Following Milligan (2019), this visualisation highlights regional sales performance briefly.

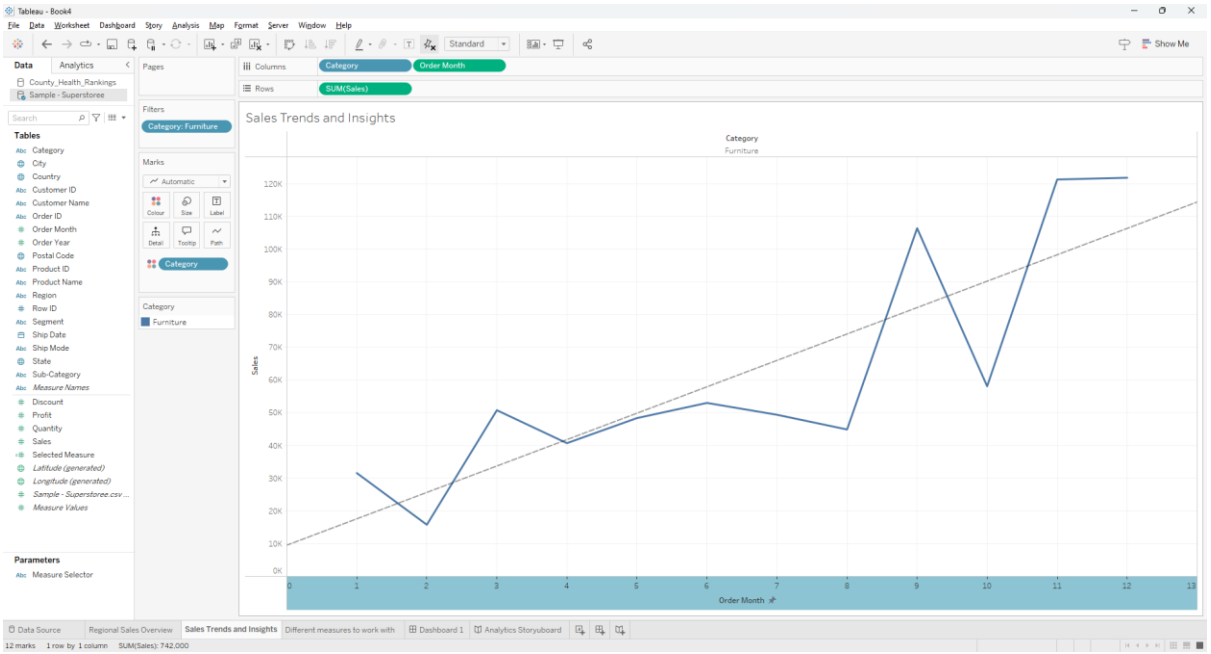


Figure 52 Sales Trends and Insights

This line chart displays monthly sales for the Furniture category, with a trend line indicating the overall growth trajectory. The chart helps identify peaks and dips in sales, providing insights into seasonal patterns or performance trends. Based on Milligan (2019), this visualisation is ideal for analysing time-based data and predicting future trends.

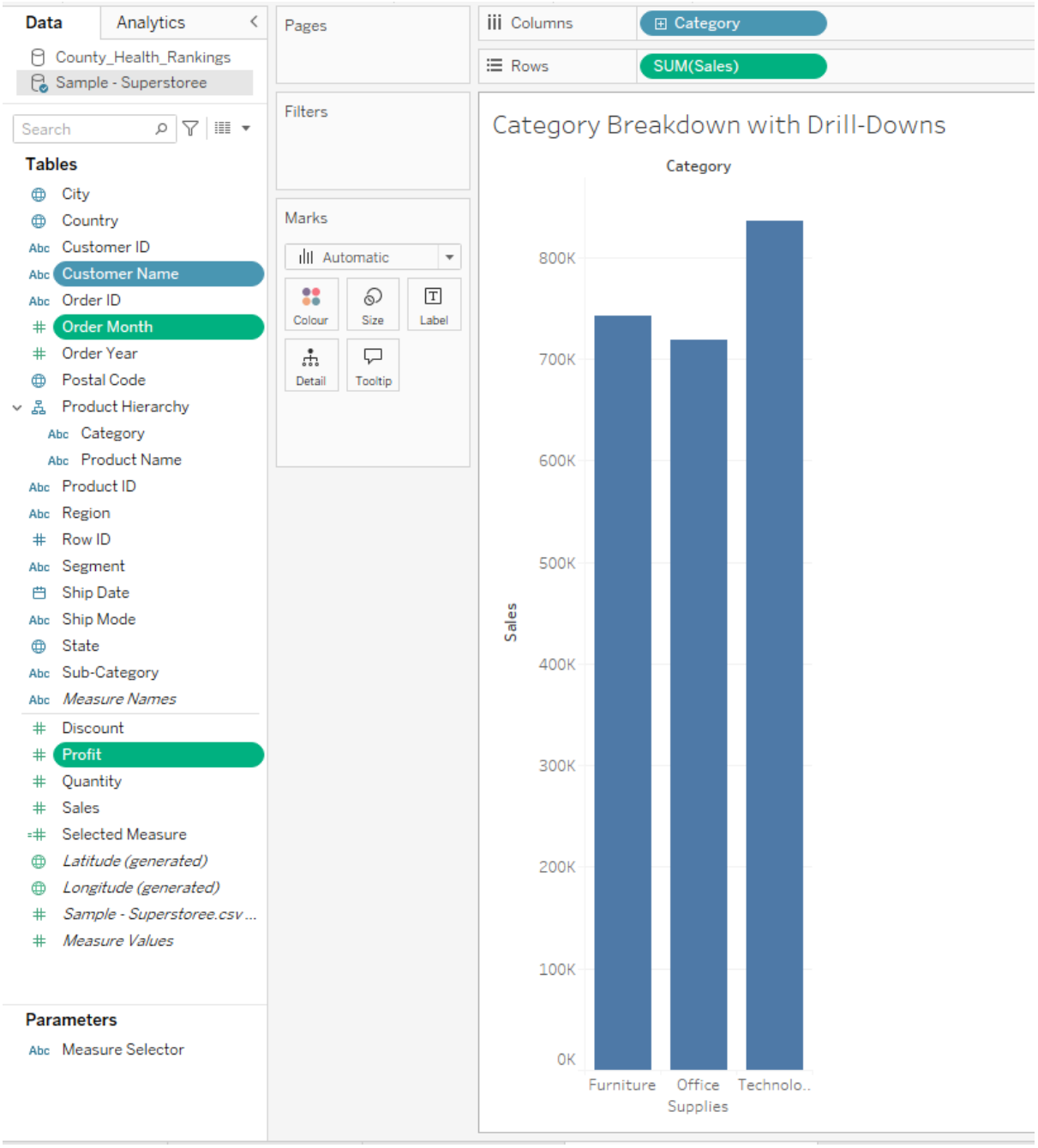


Figure 53 Category Breakdown with Drill-Downs

This bar chart visualises total sales for each product category: Furniture, Office Supplies, and Technology. The straightforward presentation highlights category-level performance, providing a clear comparison of sales volumes. The chart sets the foundation for further exploration into subcategories or products for detailed insights.

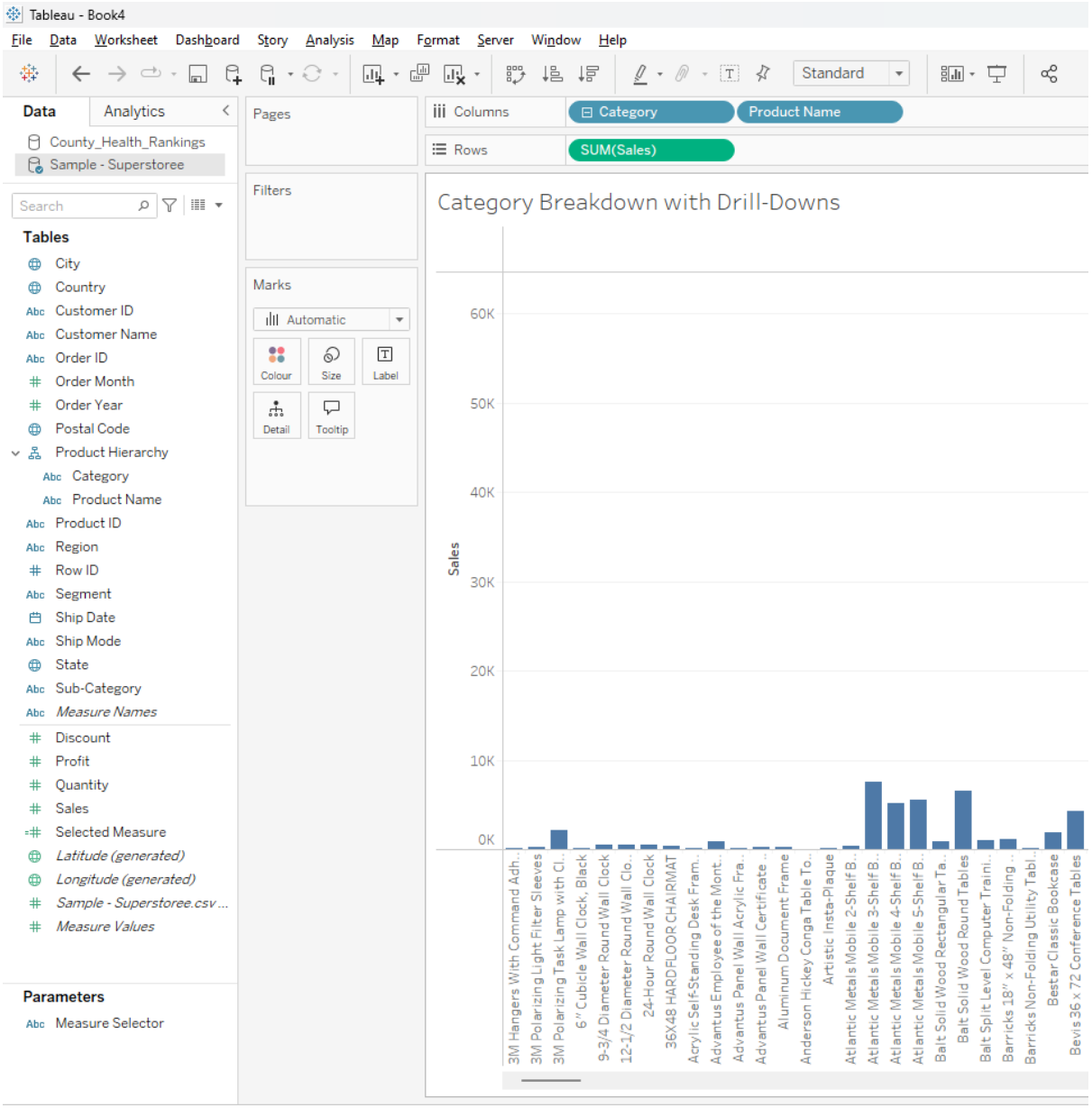


Figure 54 Product Breakdown within Categories

This detailed bar chart expands on the category-level analysis by drilling down into individual product sales within each category. While offering granular insights, the chart highlights variations in sales across products. The relationship between categories and their products is made evident, supporting a deeper understanding of sales distribution.

Both figures are part of a hierarchical exploration, with Figure 1 serving as the high-level overview and Figure 2 delving into the details for comprehensive analysis. Let me know if you need further refinements or additional insights!

5.9 Week 9

Task 9

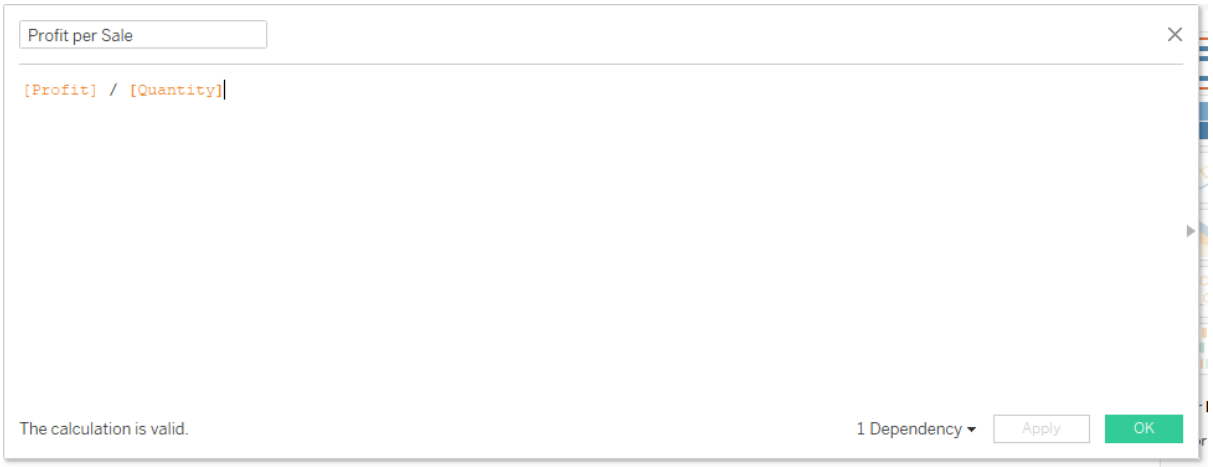


Figure 55 Profit per Sale Calculation

This figure shows the creation of a calculated field to compute 'Profit per Sale' by dividing total profit by the quantity sold. It is a foundational step for further analysis and visualisation

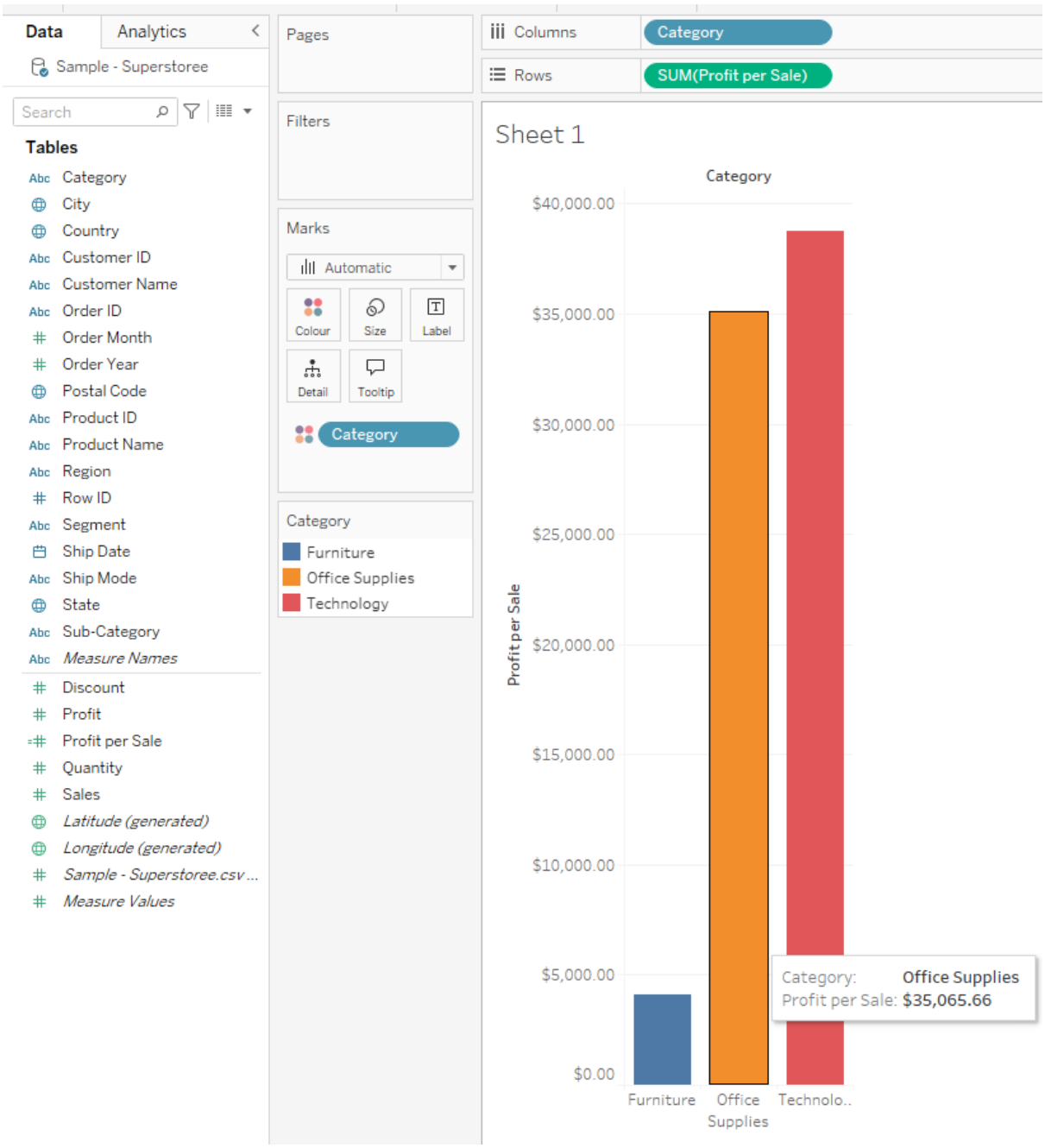


Figure 56 Profit per Sale by Category

This bar chart provides a comparison of total profit per sale across three product categories: Furniture, Office Supplies, and Technology. It clearly demonstrates that the Technology category yields the highest profit per sale, followed by Office Supplies, while Furniture shows significantly lower profitability. According to Milligan (2019), visualisations like this are crucial for identifying key performance trends, enabling data-driven decisions about

resource allocation and category focus.

Figure 57 Region & Segment Calculated Field

This screenshot displays the creation of a calculated field that combines the Region and Segment fields into a single string, separated by " - ". This calculated field serves as a foundation for future analysis and visualisation, enabling detailed insights by merging two related dimensions.

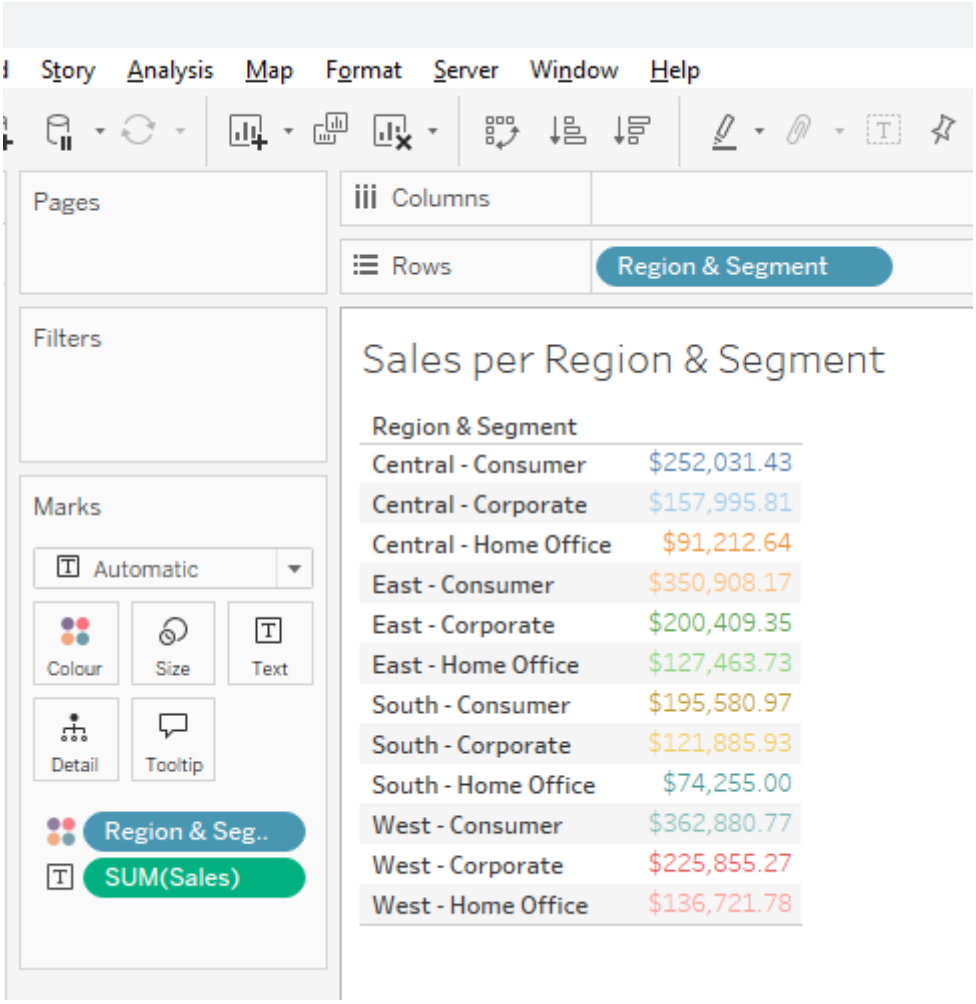


Figure 58 Sales per Region & Segment Table

This table displays total sales values grouped by a combined field, Region & Segment. Each row shows the sum of sales for specific customer segments (e.g., Consumer, Corporate, Home Office) within different regions. This provides a detailed breakdown, allowing for analysis of sales performance across regional and customer segment dimensions.

Step 2 date

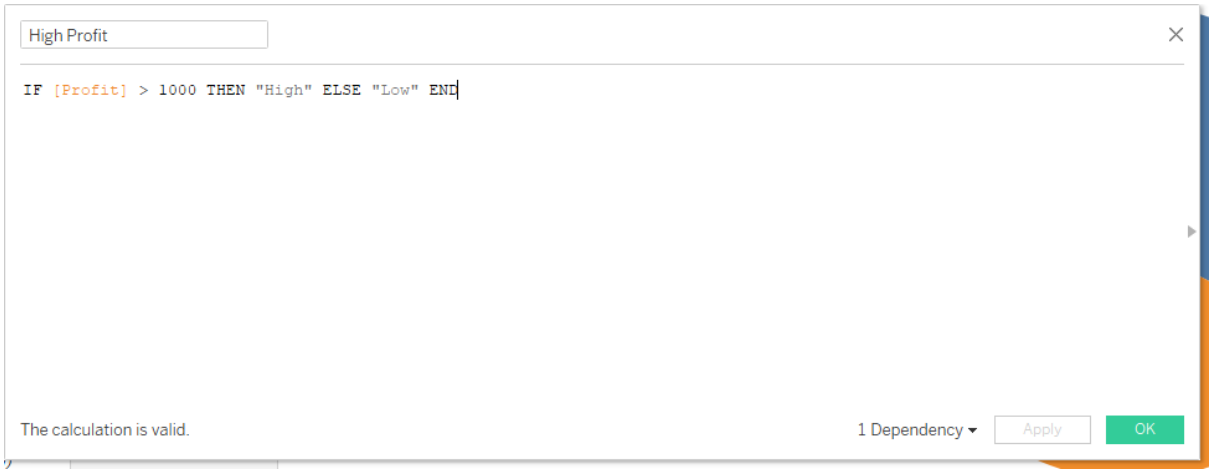


Figure 59 High Profit Logical Field

This screenshot demonstrates the creation of a calculated field named High Profit, which uses a logical condition to categorise profits as "High" if greater than 1000, and "Low" if less.

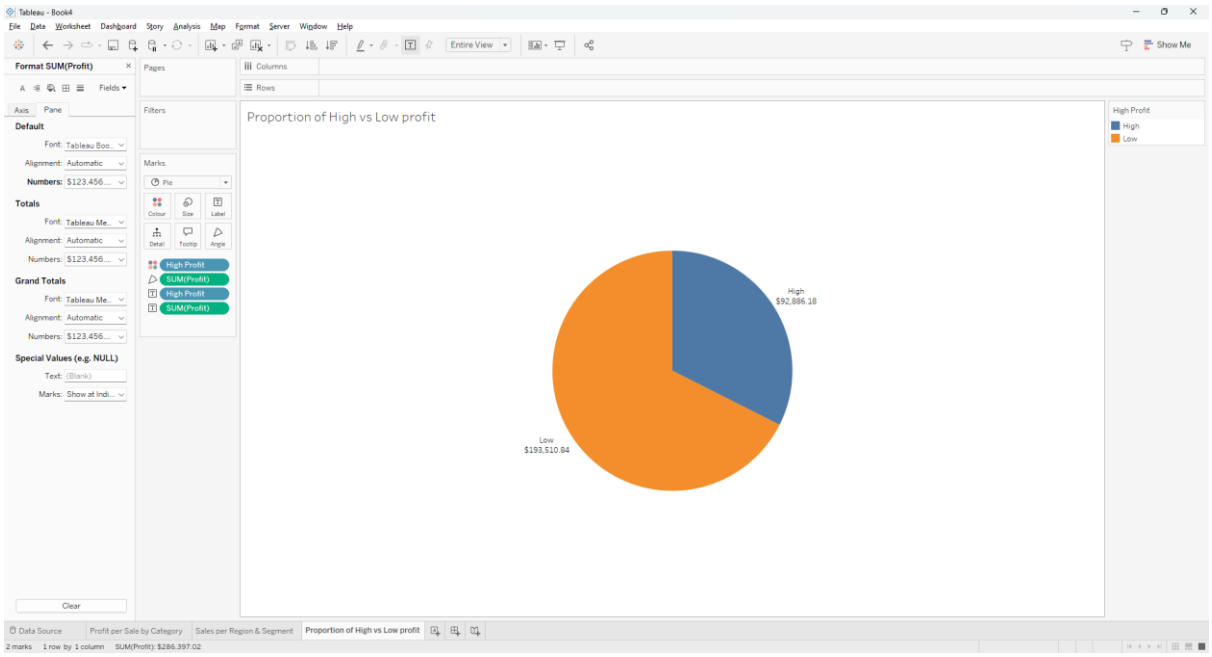


Figure 60 Proportion of High vs Low Profit

This pie chart visualises the proportion of "High" versus "Low" profit categories, based on the logical calculation where profits above 1000 are classified as "High." The chart provides a clear comparison of the two profit levels, highlighting their respective contributions to total profit.

Figure 61 Average Sales Calculated Field

This screenshot displays the creation of a calculated field named Average Sales, which computes the average value of the Sales field using the AVG function. This metric provides an overall measure of sales performance for use in further analysis or visualisations.

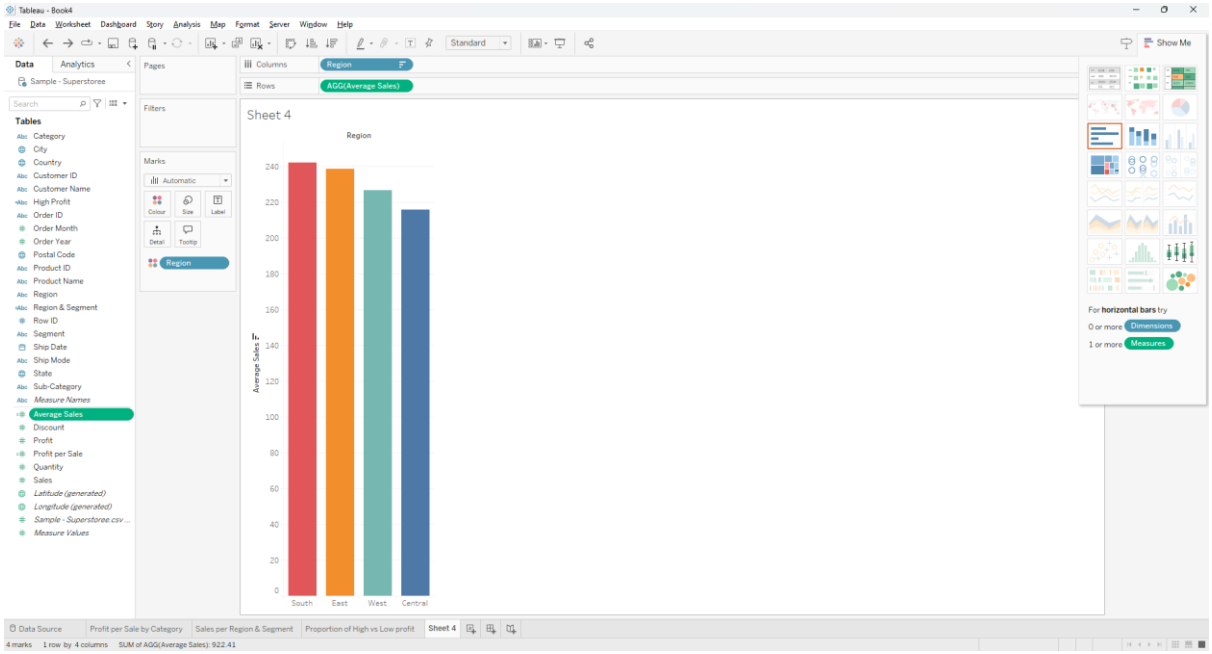


Figure 62 Average Sales by Region

This bar chart visualises the average sales across different regions (South, East, West, and Central). It highlights regional differences in sales performance, providing insights into which regions have higher or lower average sales values.

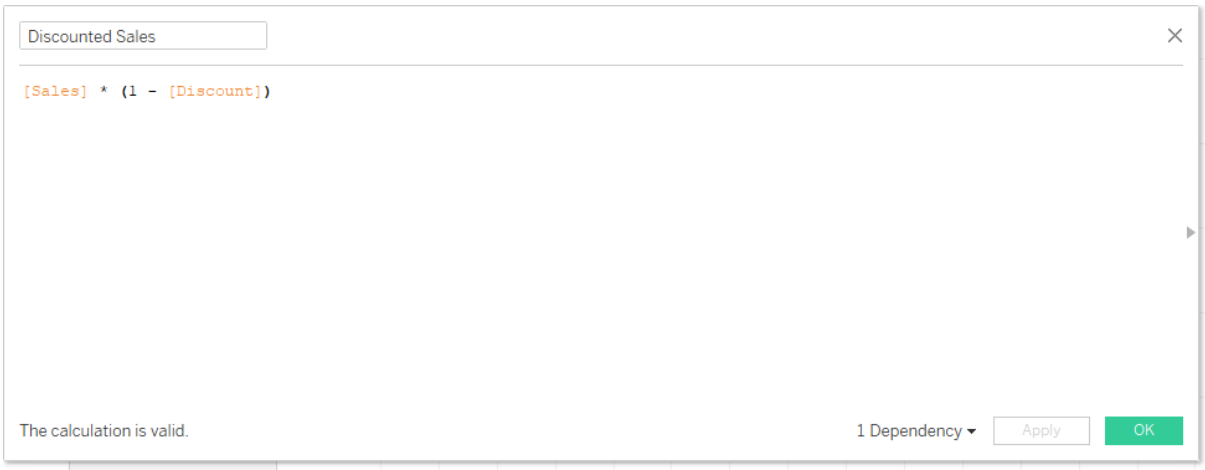


Figure 63 Discounted Sales Calculated Field

This screenshot illustrates the creation of a calculated field named Discounted Sales, which computes the sales after applying discounts. The formula multiplies the original Sales by (1 - Discount), accounting for the percentage discount applied to each sale. This field is useful for evaluating the impact of discounts on revenue.

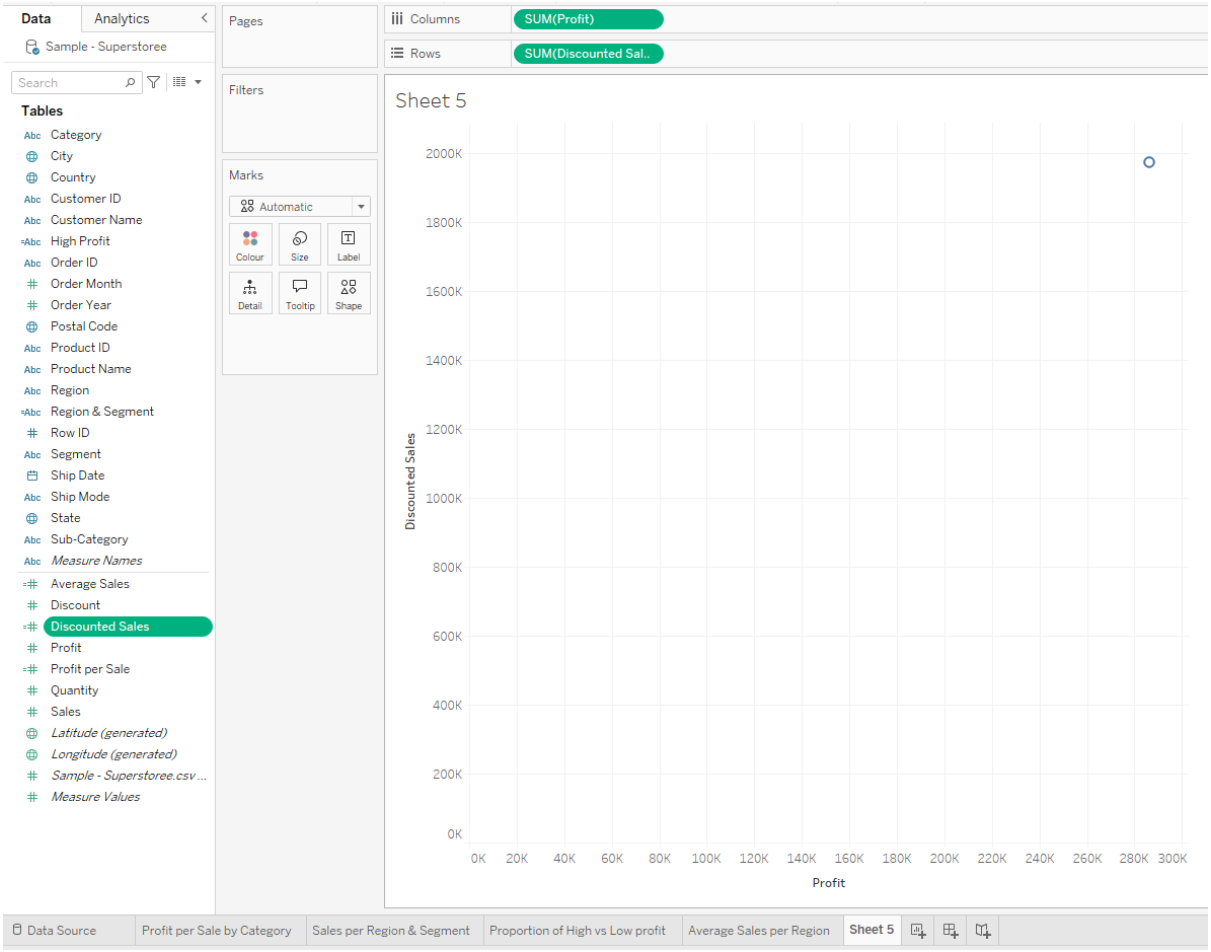


Figure 64 Discounted Sales vs Profit

This scatter plot visualises the relationship between Discounted Sales and Profit. Each point represents a specific data entry, allowing for the analysis of how discounted sales correlate with overall profitability. This helps identify whether discounts are driving profitable outcomes or negatively impacting profits.

Figure 65 Customer Region Calculated Field

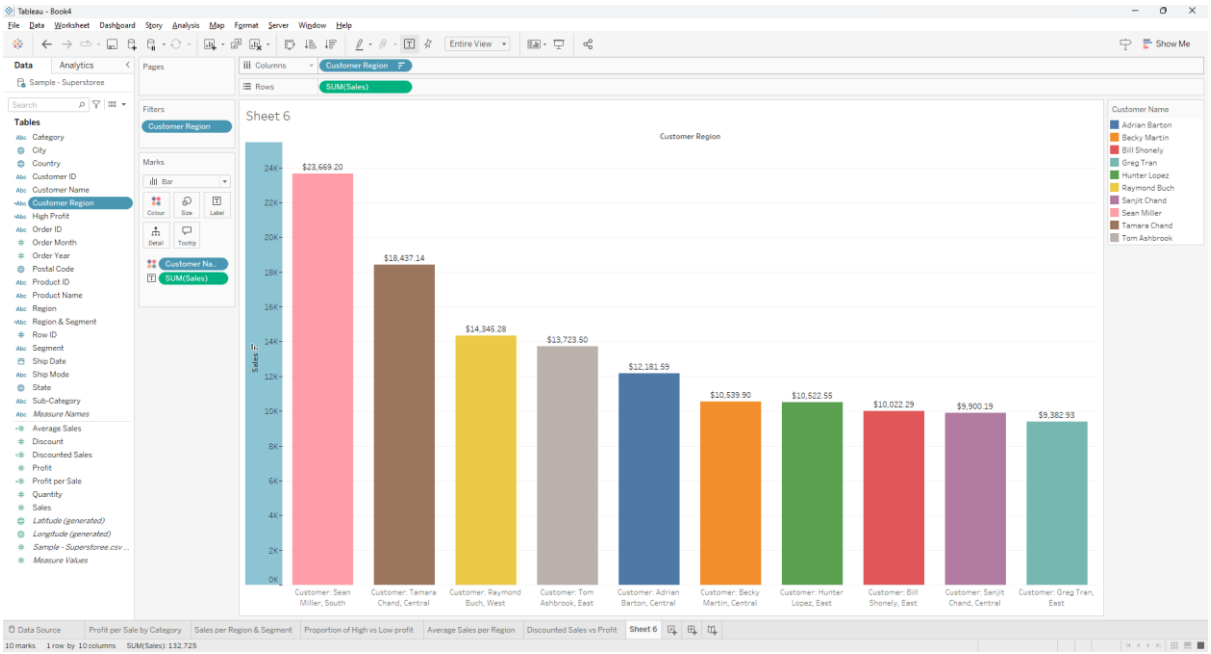


Figure 66 Sales by Customer Region

This bar chart displays the total sales associated with individual customers, categorised by their regions. The Customer Region calculated field combines customer names and regions to provide a detailed breakdown of sales performance for each customer-location pair. This visualisation helps identify high-value customers across different regions.

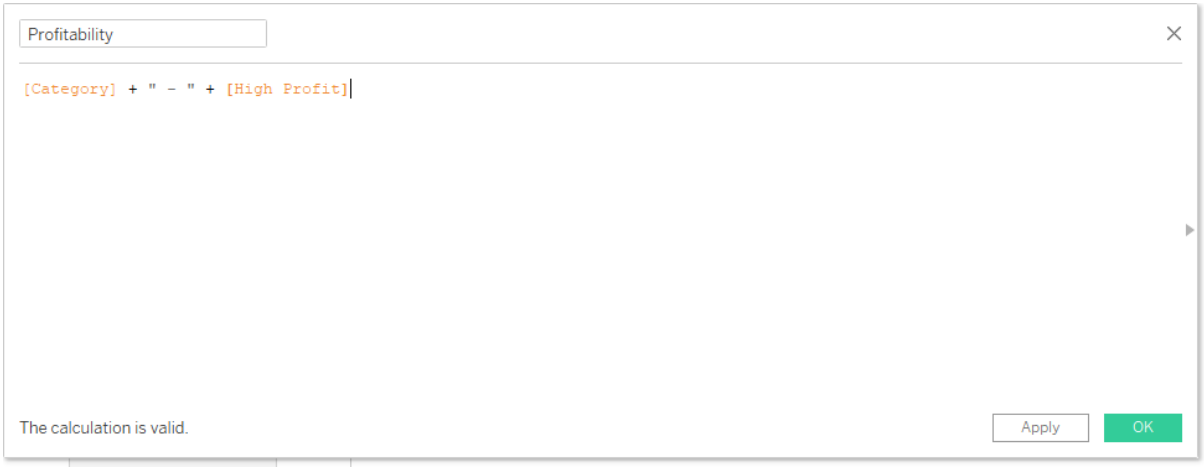


Figure 67 Profitability Calculated Field

This screenshot shows the creation of a calculated field named Profitability, which combines the Category and High Profit fields into a single string separated by " - ". This field is useful for grouping and analysing product categories based on their profit classifications (High or Low).

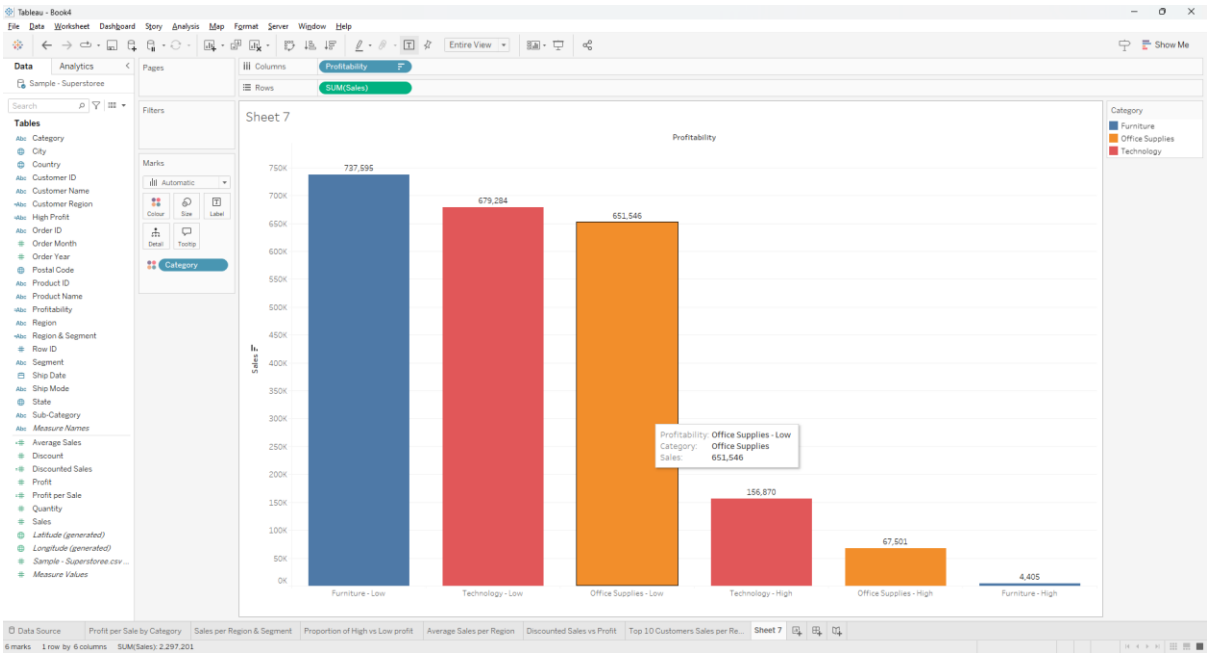


Figure 68 Profitability by Category and Sales

This bar chart visualises total sales segmented by the Profitability field, which combines product categories (Furniture, Office Supplies, and Technology) with profit classifications (High or Low). It highlights how each category performs in terms of sales volume for both high and low profitability, providing insights into the relationship between profit margins and sales.

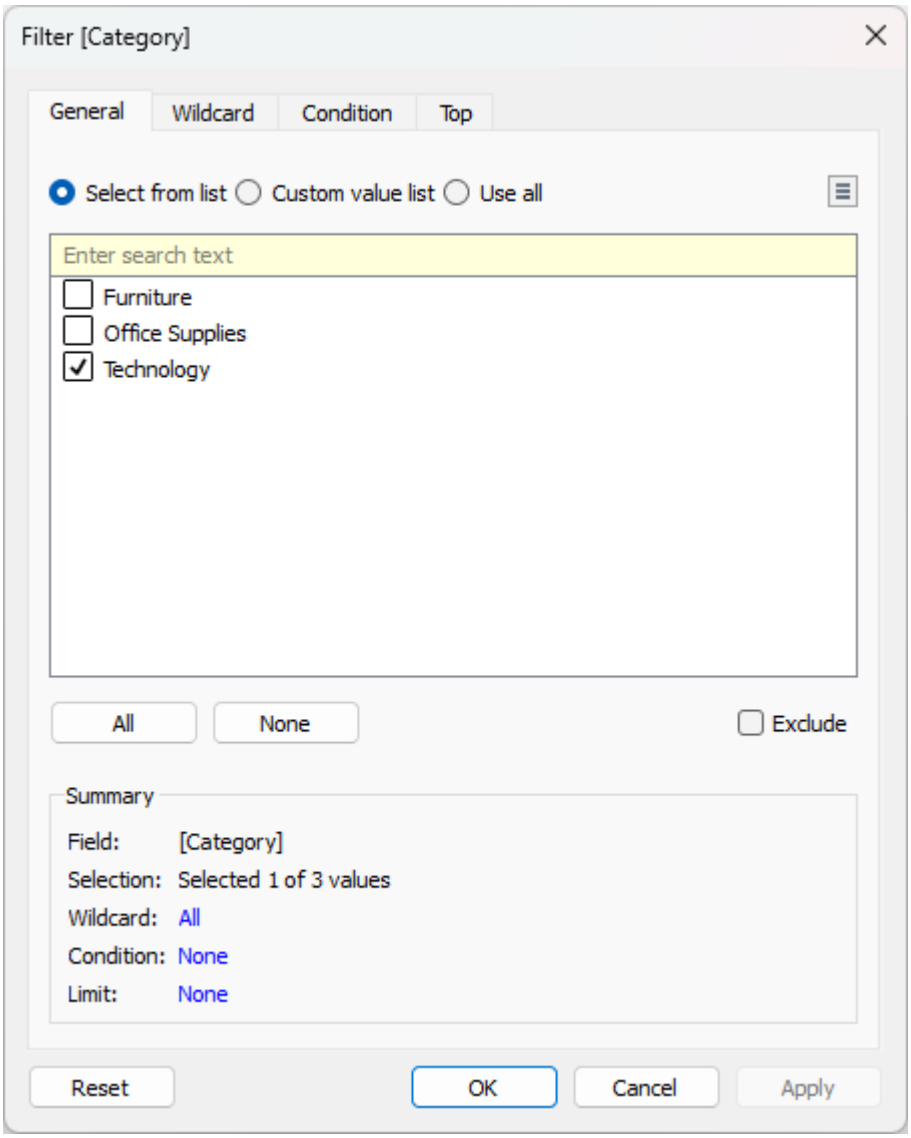


Figure 69 Category Filter Selection

This screenshot displays the configuration of a filter for the Category field, where only the "Technology" category has been selected. Filters like this allow for targeted analysis by isolating specific categories, enabling focused insights and cleaner visualisations.

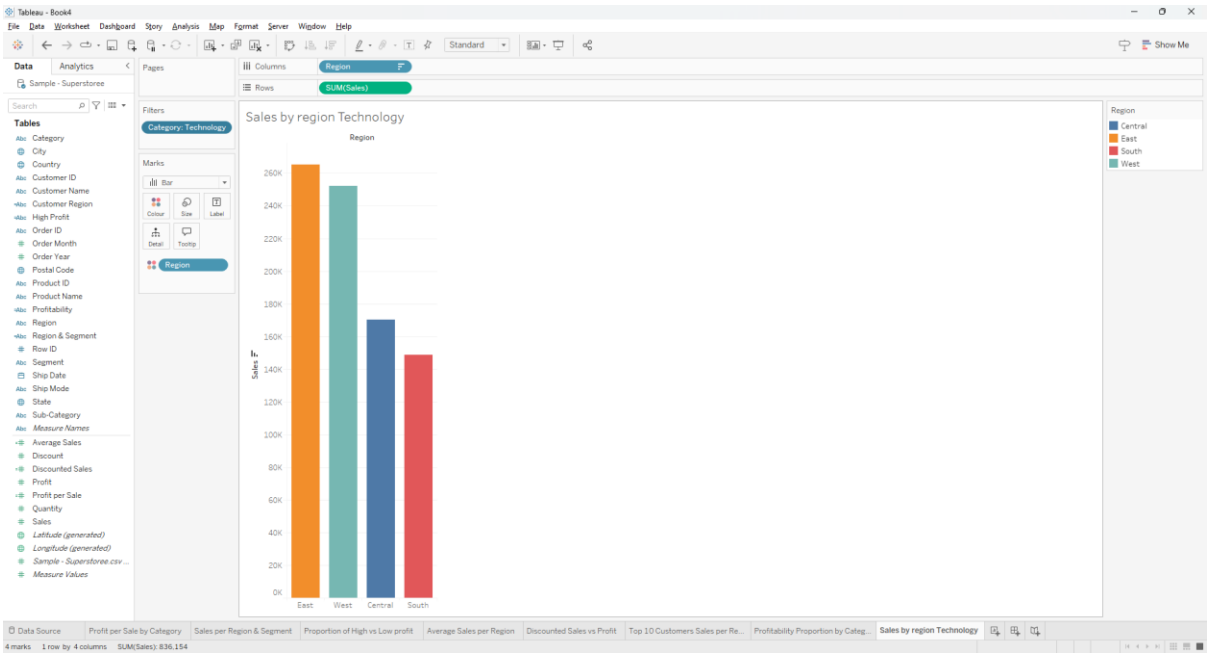


Figure 70 Sales by Region for Technology Category

This bar chart presents total sales for the Technology category segmented by region (East, West, Central, and South). The visualisation focuses solely on the Technology category, as specified by the filter applied. According to Milligan (2019), using filters effectively helps narrow down large datasets, enabling analysts to derive focused insights and make more informed decisions based on specific criteria. This approach ensures clarity in identifying regional performance within a single product category.

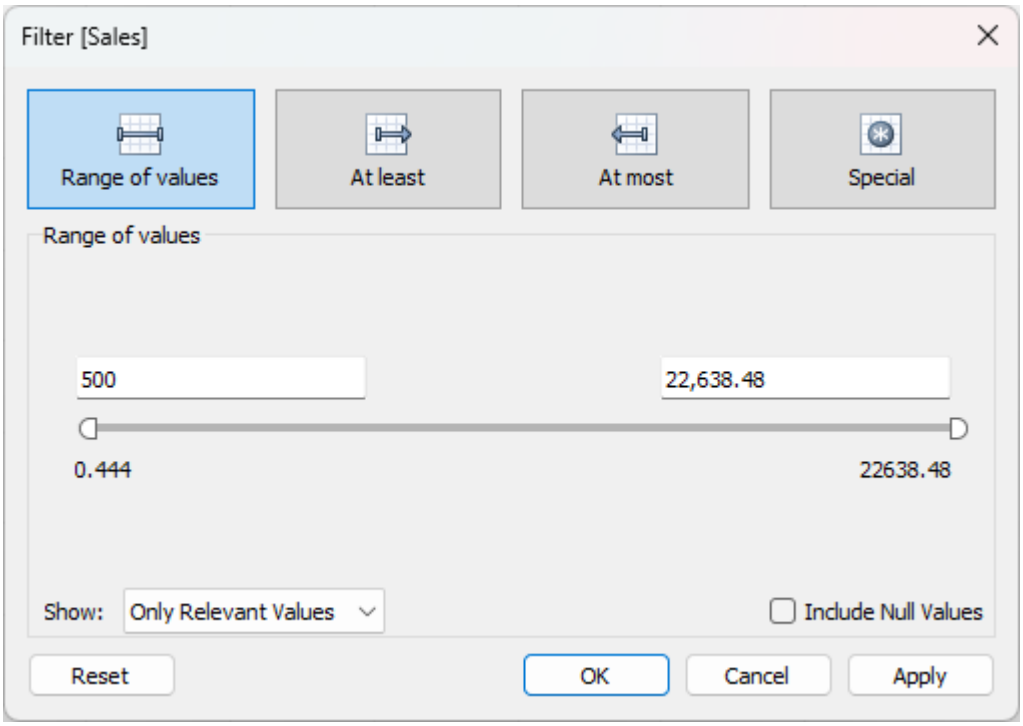


Figure 71 Sales Range Filter

This screenshot shows a filter applied to the Sales field, specifying a range between $500 and $22,638.48. Using a range filter allows analysts to focus on transactions within a specific sales threshold, eliminating smaller or outlier values that may skew the results. This technique refines the dataset for targeted analysis and better insights.

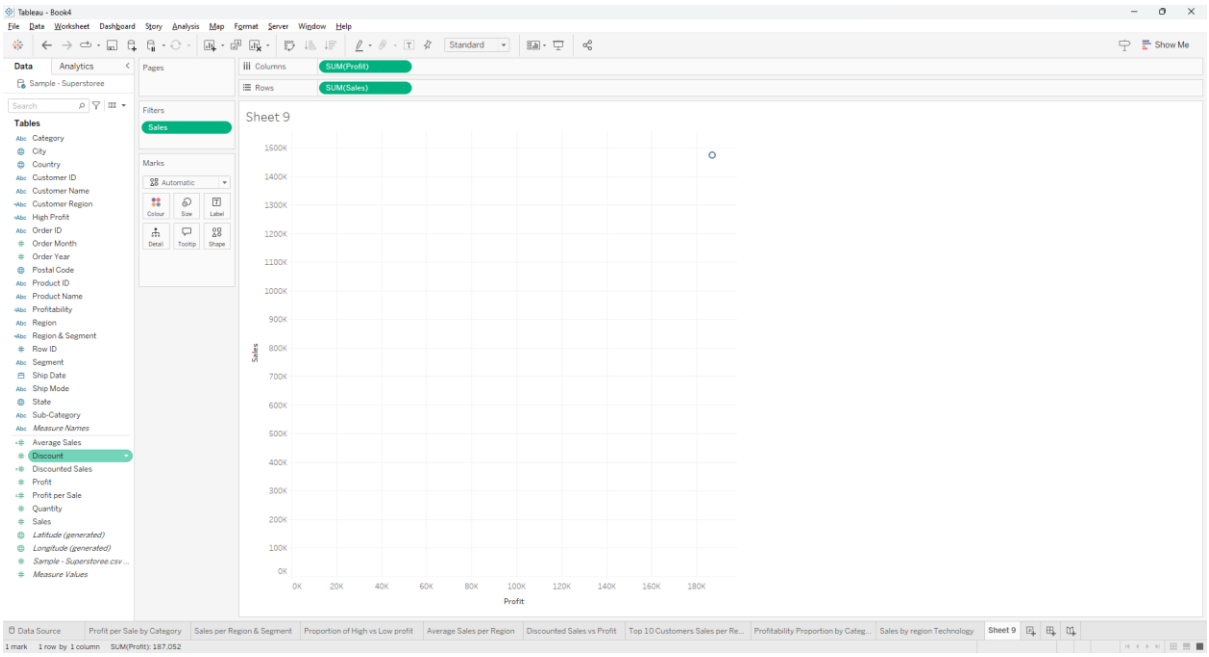


Figure 72 Scatter Plot of Profit vs Sales with Sales Filter Applied

This scatter plot visualises the relationship between total profit (SUM(Profit)) and total sales (SUM(Sales)) while applying a filter to include only sales greater than £500. The plot allows for the identification of trends or correlations between these two variables, focusing on transactions with significant sales values.

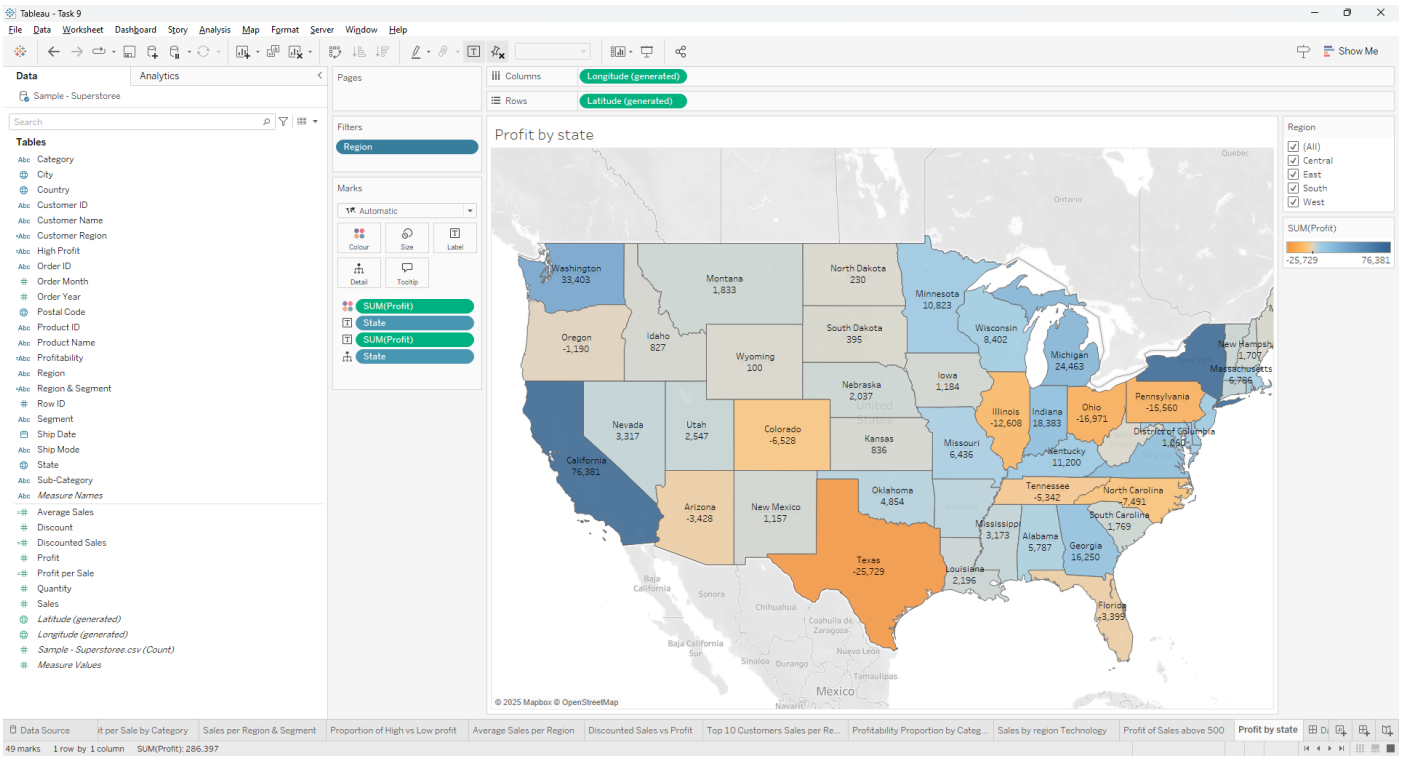


Figure 73 Profit by State Map

This filled map visualises the profit distribution across different U.S. states. Each state is shaded based on its profit value, with varying colours representing profit levels. States with negative profits are highlighted in orange, indicating losses, while states with higher profits are shaded in darker blue. The map provides a geographical overview of profitability, allowing for easy identification of high-performing and underperforming areas.

5.10 Week 10

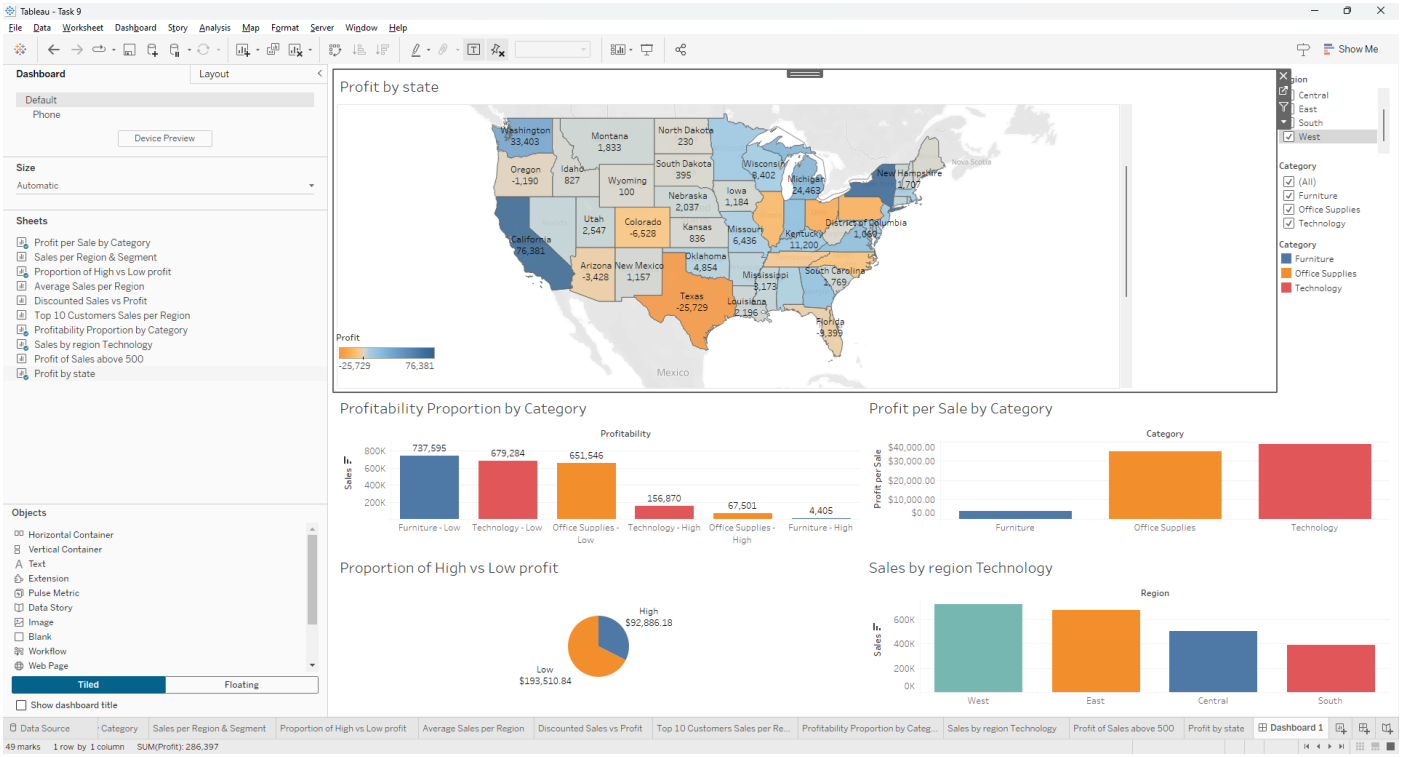


Figure 74 Sales and Profit Analysis Dashboard

This dashboard combines multiple visualisations to provide a comprehensive analysis of sales and profitability across regions, categories, and profitability levels. It aligns with Milligan's (2019) observation that "effective visualisations distil complex datasets into actionable insights," enabling decision-makers to identify high-performing regions and categories, focus on improving low-profit areas, and optimise resource allocation. By integrating interactive filters for regions and categories, the dashboard encourages exploration and supports strategic business planning.

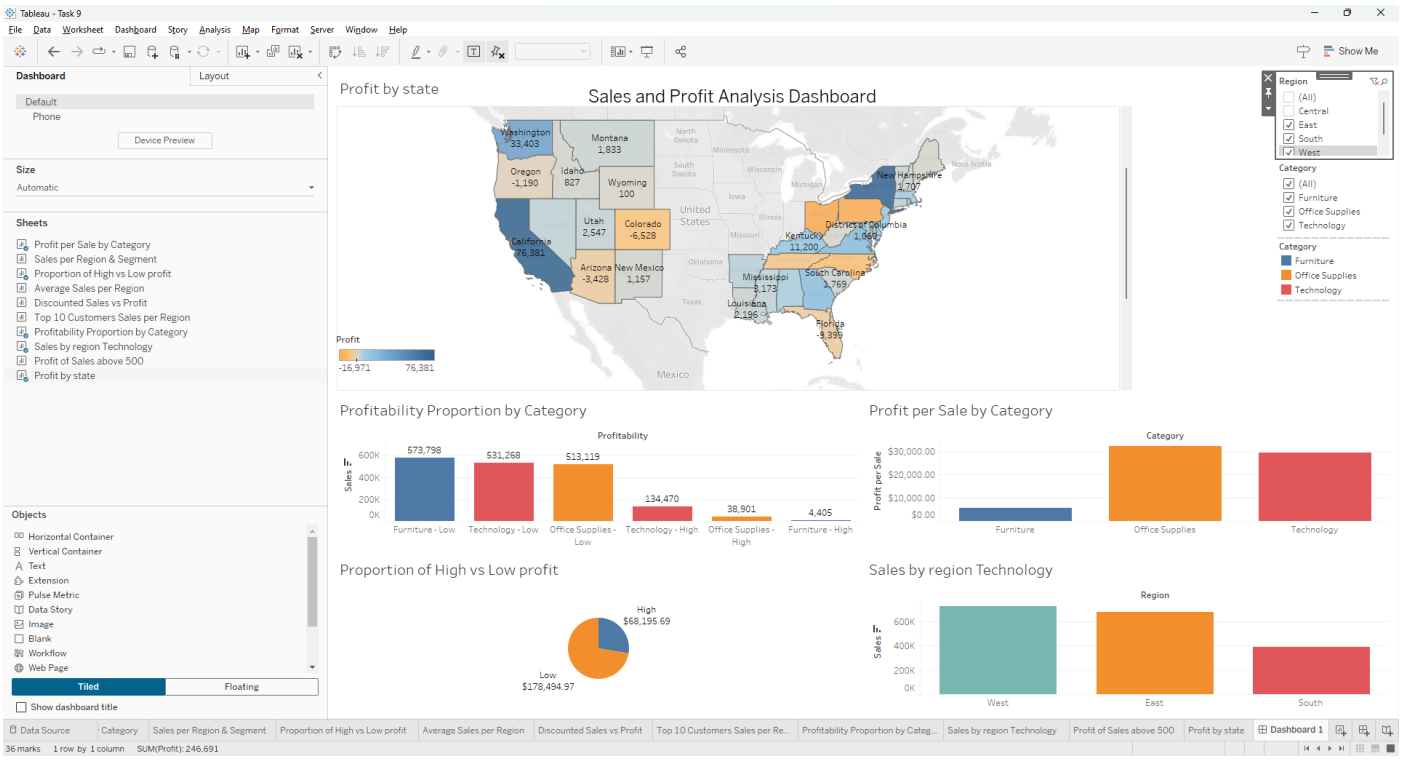


Figure 75 Sales and Profit Analysis Dashboard

This interactive dashboard provides a dynamic exploration of sales and profit data. By unticking "Central" from the region filter, the dashboard adjusts all related visualisations in real-time, reflecting updated insights for the remaining regions (East, South, and West). This interconnected design ensures that users can quickly adapt their analysis to focus on specific dimensions or scenarios.

Milligan (2019) underscores that "the true value of a dashboard lies in its ability to respond to user input, creating a dialogue with the data." This dashboard embodies that principle, enabling seamless interactivity and fostering deeper engagement with the insights.

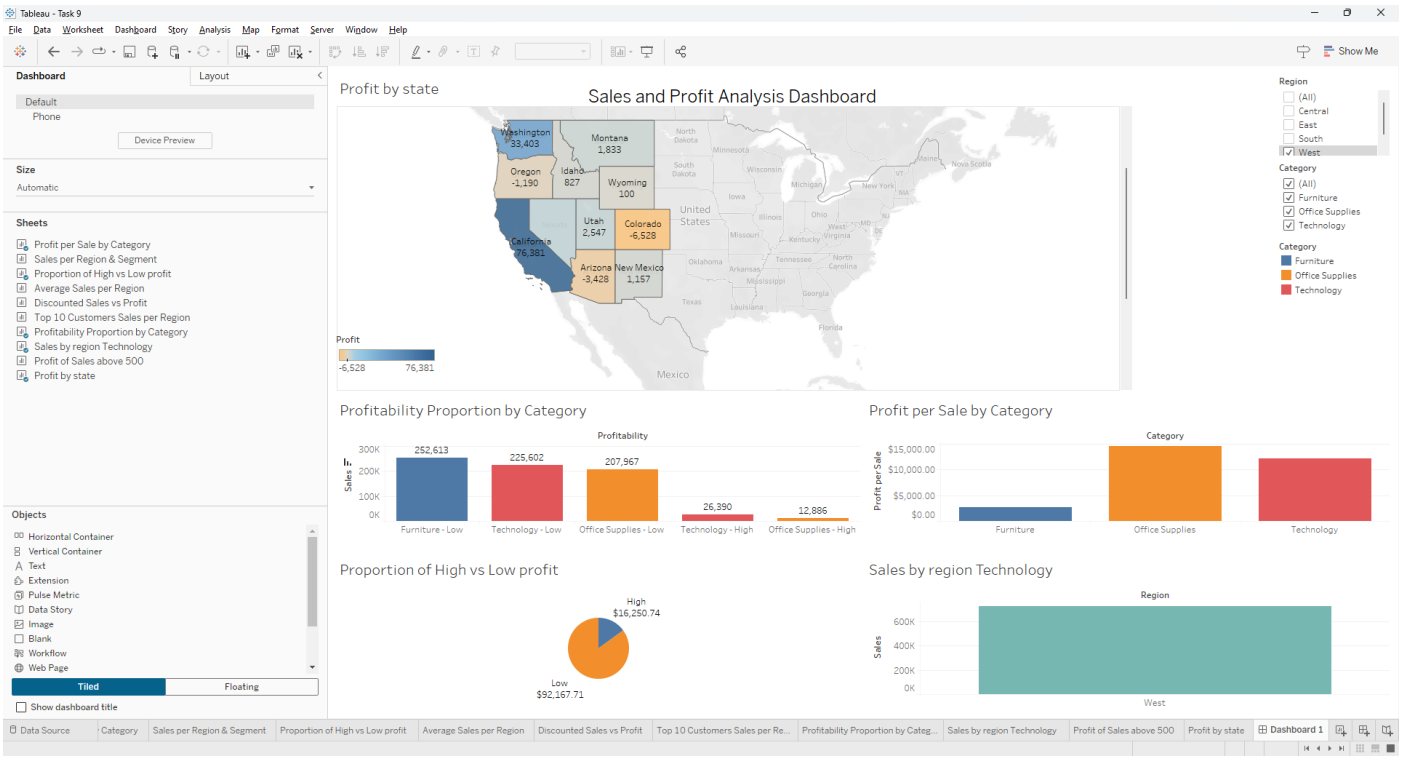


Figure 76 Sales and Profit Analysis Dashboard

This dashboard demonstrates how interconnected visualisations provide real-time insights as users interact with filters. In this case, only the "West" region is selected, and all visualisations dynamically update to reflect the filtered data. This allows users to drill down into specific geographic areas and categories, making it easier to identify trends and anomalies.

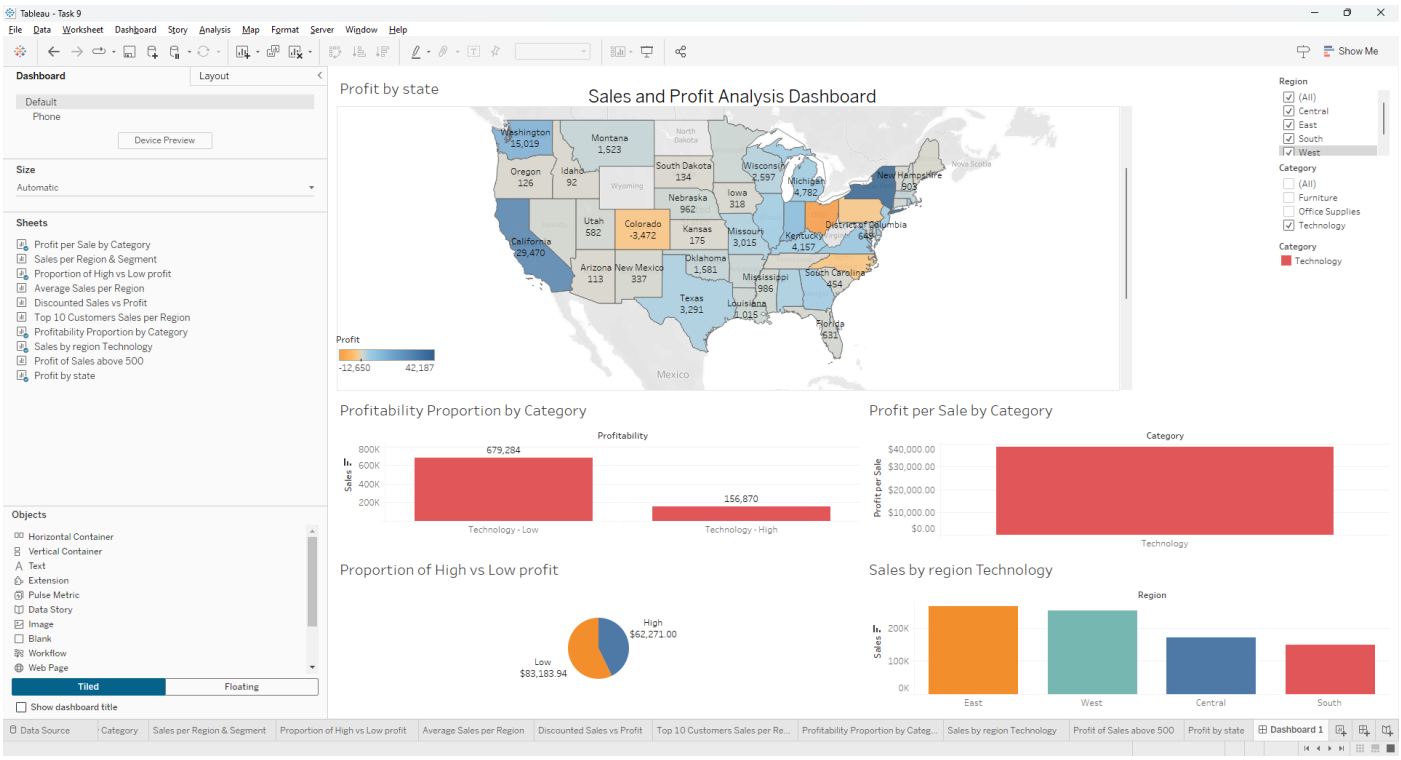


Figure 77 Sales and Profit Analysis Dashboard with Technology Focus

This dashboard dynamically adapts to user inputs, showcasing insights specific to selected categories and regions. In this example, the "Technology" category was selected, and all charts updated to reflect only the relevant data. The interactivity ensures that users can drill down into specific areas of interest, such as focusing solely on the Technology category's sales and profit trends across regions.

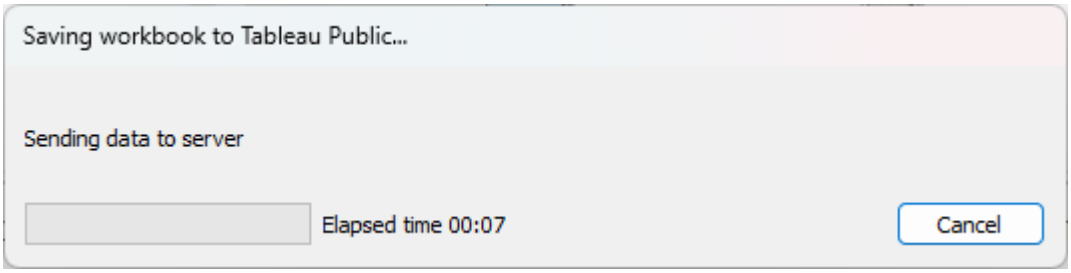


Figure 78 Uploading Workbook to Tableau Public

This screen capture shows the process of saving a workbook to Tableau Public, where data is being sent to the server, enabling online sharing and interaction with the visualised analyses.

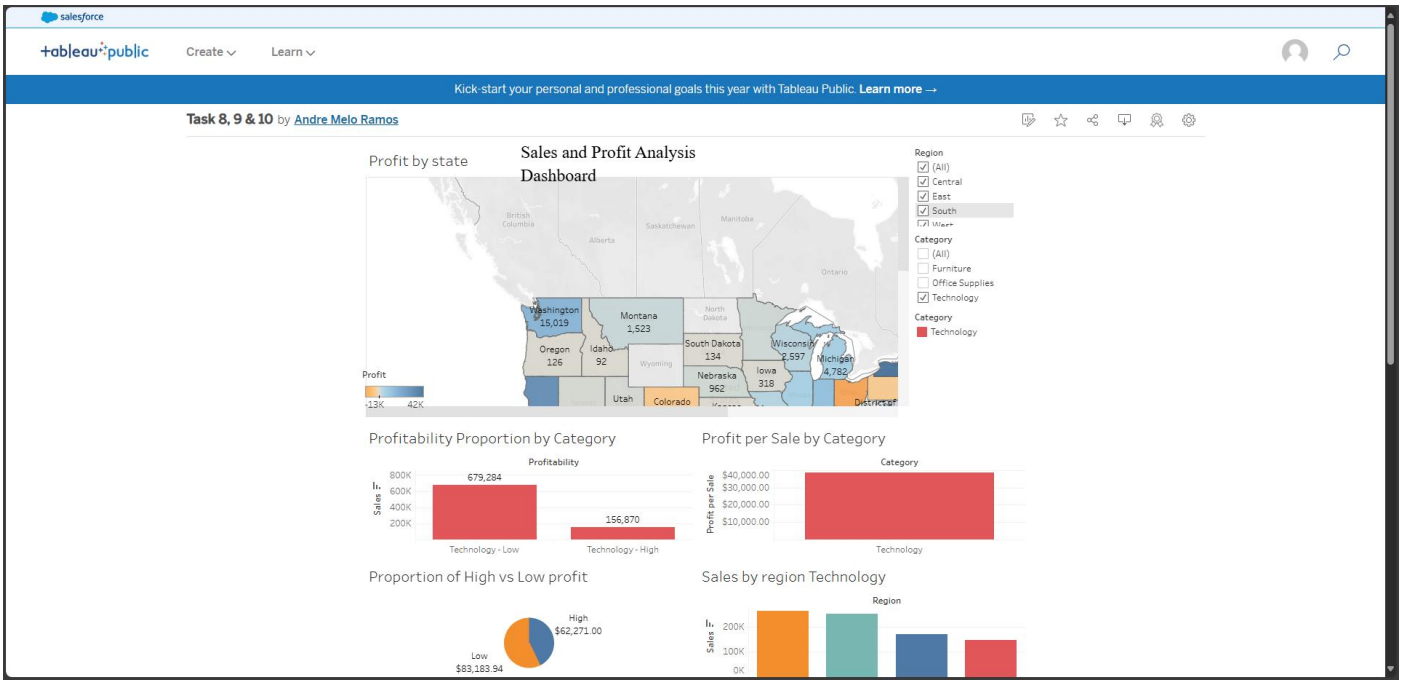


Figure 79 Sales and Profit Analysis Dashboard on Tableau Public

This image showcases a comprehensive dashboard hosted on Tableau Public, displaying various data visualisations

such as 'Profit by State', 'Profitability Proportion by Category', and 'Sales by Region Technology'. This dashboard allows for interactive exploration of data, providing insights into sales performance and profitability across different categories and regions, directly accessible through Tableau Public.

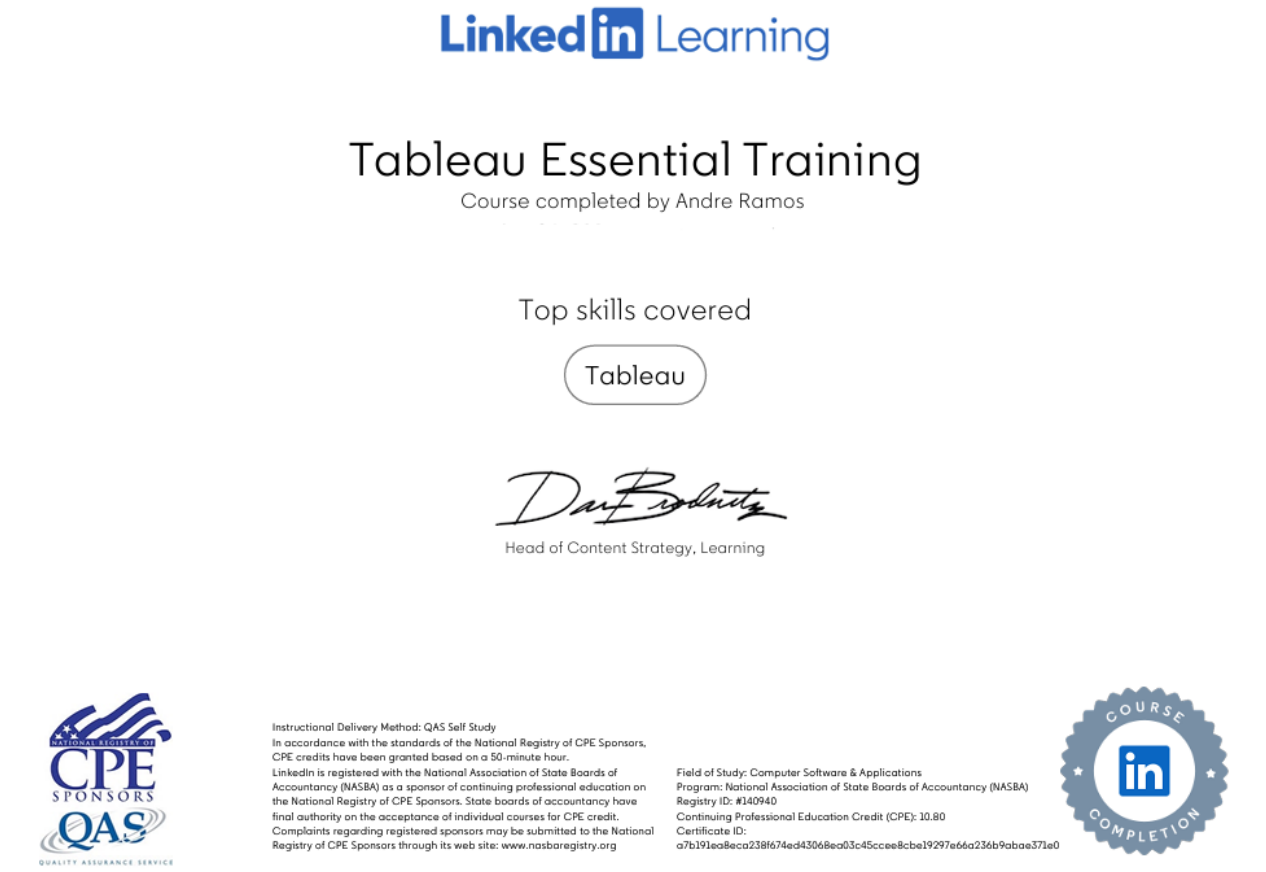


Figure 80 Tableau Essential Training Completion Certificate

This image shows a completion certificate from LinkedIn Learning for the "Tableau Essential Training" course completed by me. The certificate highlights that the course covers fundamental skills in Tableau, a key tool for data visualisation and analysis. This acknowledgment certifies the acquisition of vital competencies in utilising Tableau for practical data handling and business intelligence tasks.