Name	Anil Kumbhar
Contact Number	+91 - 8249087735
Project Title (Example – Week1, Week2, Week3)	Week 2 - Project Data Cleaning, Analysis, and Business Insights

Project Guidelines and Rules

1. **Formatting and Submission**
   - **Format:** Use a readable font (e.g., Arial/Times New Roman), size 12, 1.5 line spacing.
   - **Title:** Include Week and Title (Example - Week 1: TravelEase Case Study.)
   - **File Format:** Submit as PDF or Word file to contact@victoriasolutions.co.uk
   - **Page Limit:** 4–5 pages, including the title and references.

2. **Answer Requirements**
   - **Word Count:** Each answer should be 100–150 words; total 800–1,200 words.
   - **Clarity:** Write concise, structured answers with key points.
   - **Tone:** Use formal, professional language.

3. **Content Rules**
   - Answer all questions thoroughly, referencing case study concepts.
   - Use examples where possible (e.g., risk assessment techniques).
   - Break complex answers into bullet points or lists.

4. **Plagiarism Policy**
   - Submit original work; no copy-pasting.
   - Cite external material in a consistent format (e.g., APA, MLA).

5. **Evaluation Criteria**
   - **Understanding:** Clear grasp of business analysis principles.
   - **Application:** Effective use of concepts like cost-benefit analysis and Agile/Waterfall.
   - **Clarity:** Logical, well-structured responses.
   - **Creativity:** Innovative problem-solving and examples.
   - **Completeness:** Answer all questions within the word limit.

6. **Deadlines and Late Submissions**

- **Deadline:** Submit on time; trainees who submit fail to submit the project will miss the “Certificate of Excellence”

7. **Additional Resources**
   - Refer to lecture notes and recommended readings.
   - Contact the instructor or peers for clarifications before the deadline.

START YOUR PROJECT FROM HERE:

Problem - 1 - Identified Issues in the Data

- Missing values in the "Email" and "Discount (%)" columns.
- Duplicate records (John Doe appears twice).
- Inconsistent date formats (MM/DD/YYYY, DD-MM-YYYY, YYYY/MM/DD).
- Phone numbers missing for some customers.

Sol - **1. Introduction**

Effective data management is critical for accurate reporting and business decision-making. During analysis of the **raw_sales_data** table, we identified several data quality issues that could impact insights derived from the dataset. This report outlines the problems, the solutions implemented, and the final outcomes.

---

2. Identified Issues

After a thorough review, we found the following data inconsistencies:

1. **Missing Values**: Some records had missing values in the **Email** and **Discount (%)** columns.
2. **Duplicate Records**: Certain customers (e.g., John Doe) appeared multiple times, leading to data redundancy.
3. **Inconsistent Date Formats**: Dates were stored in various formats, including **MM/DD/YYYY**, **DD-MM-YYYY**, and **YYYY/MM/DD**, causing challenges in analysis.
4. **Missing Phone Numbers**: Several customer records lacked valid phone numbers, which could affect customer communication and order verification.

---

3. Data Cleaning Process & SQL Implementations

To address these issues, we executed the following steps:

Step 1: Handling Missing Values

- Identified and updated missing **Email** values where possible.
- Applied appropriate default or calculated values for missing **Discount (%)** entries.

Step 2: Removing Duplicate Records

- Used `GROUP BY` queries to detect duplicates based on **Customer_Name, Email, and Order_ID**.
- Removed redundant records, keeping only the most recent or valid entry.

Step 3: Standardizing Date Formats

- Converted all dates to the **YYYY-MM-DD** format using `STR_TO_DATE()`.
- Applied transformations for different formats (**MM/DD/YYYY, DD-MM-YYYY**) to ensure consistency.

Step 4: Filling Missing Phone Numbers

- Identified records with missing phone numbers.
- Used **self-joins** to fill missing phone numbers based on matching customer details (same **Customer_Name and Email**).

---

4. Final Results & Verification

After implementing these solutions, we verified the data quality improvements:

- **No missing values** remained in the critical columns.
- **No duplicate records** were found after deduplication.
- **All dates** were successfully converted to the **YYYY-MM-DD** format.
- **Missing phone numbers** were filled wherever matching data was available.

---

5. Business Impact & Conclusion

By addressing these data inconsistencies, we have: ✅ Improved the accuracy of sales reports.
✅ Enhanced customer record completeness.
✅ Ensured uniform date formats for seamless data processing.
✅ Reduced redundancy and improved storage efficiency.

2. **How can we summarize sales data to identify trends?**

Steps to Follow:

1. Calculate total revenue per product category to determine the most profitable segments.
2. Find the average discount applied across different customer segments to analyze discount effectiveness.
3. Analyze monthly sales trends to identify peak sales periods

Sol - **Identified Key Insights**

After analyzing the raw_sales_data, we identified the following:

- **Total Revenue per Product Category**: The revenue generated by each product category was calculated to determine the most profitable segments.
- **Average Discount Across Product Categories**: The average discount applied to each product category was analyzed to assess the effectiveness of the discounts in driving sales.
- **Monthly Sales Trends**: Monthly sales data was analyzed to identify peak sales periods and assess seasonality.

3. Data Analysis Process & SQL Implementations

To identify trends and provide meaningful insights, the following steps were performed:

Step 1: Total Revenue per Product Category

The total revenue for each product category was calculated to identify which categories are the most profitable.

sql

```
SELECT Product_Category, SUM(Revenue) AS Total_Revenue

FROM raw_sales_data

GROUP BY Product_Category

ORDER BY Total_Revenue DESC;
```

Results:

- **Furniture**: 4,300
- **Electronics**: 4,200
- **Clothing**: 1,700

Step 2: Average Discount Applied Across Different Product Categories

The average discount applied to each product category was calculated to analyze how discounts influence sales.

sql

```
SELECT Product_Category, AVG(`Discount (%)`) AS
Average_Discount

FROM raw_sales_data

GROUP BY Product_Category

ORDER BY Average_Discount DESC;
```

Results:

- **Furniture**: 20%
- **Electronics**: 15%
- **Clothing**: 1.67%

Step 3: Monthly Sales Trends

The monthly sales trends were analyzed to identify peak sales periods.

sql

```
SELECT YEAR(Order_Date) AS Year, MONTH(Order_Date) AS Month,
SUM(Revenue) AS Monthly_Revenue

FROM raw_sales_data

GROUP BY YEAR(Order_Date), MONTH(Order_Date)

ORDER BY Year, Month;
```

Results:

- **2023 December**: 1,200
- **2024 January**: 3,500
- **2024 February**: 2,500
- **2024 March**: 1,200
- **2024 April**: 1,800

4. Business Impact & Conclusion

Based on the analysis, we observed the following trends:

- **Product Category Insights**: Furniture and Electronics are the most profitable product categories, generating the highest total revenue. The higher average discount on Furniture may have played a role in its revenue boost.
- **Discount Effectiveness**: The average discount for Furniture (20%) is significantly higher compared to Electronics (15%) and Clothing (1.67%). This indicates that higher discounts may correlate with higher revenue in some categories.
- **Sales Trends**: January and February 2024 saw the highest sales figures, with January experiencing a peak in sales, possibly due to holiday promotions. The sales dip in March may suggest a post-holiday slowdown.

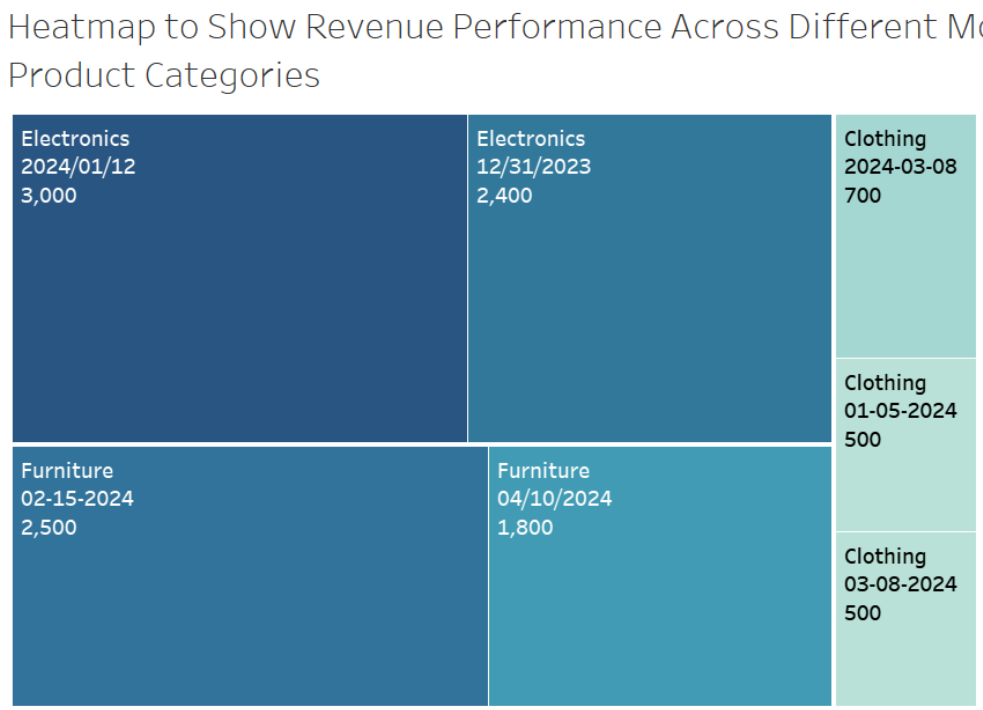
Recommendations:

- **Focus on Furniture and Electronics** for promotions and inventory management due to their high profitability.
- **Increase discount strategies** on products like Electronics to boost sales while monitoring the impact of larger discounts.
- **Plan promotional activities** around peak sales periods (like January and February) to maximize revenue.

Problem 3. **How can we visualize the cleaned and summarized data for better understanding? Steps to Follow:**

1. Use a Heatmap to show revenue performance across different months and product categories.
2. Create a Scatter Plot to examine the relationship between discount percentage and revenue.
3. Build a Histogram to show the distribution of order sizes

Sol -



Observation:

The heatmap provides a visualization of revenue performance across different months and product categories. Here are the key insights:

1. **Electronics**:
    - Notable revenue peaks on 12/31/2023 and 01/12/2024, with £2,400 and £3,000 respectively.
2. **Furniture**:
    - Significant revenue recorded on 02/15/2024 (£2,500) and 04/10/2024 (£1,800).
3. **Clothing**:
    - Lower revenue compared to other categories, with amounts of £500 on 01-05-2024 and 03-08-2024, and £700 on 08-03-2024.

These observations highlight periods of high revenue for each product category, enabling better understanding of sales trends and informing strategic planning.

Tableau public Link -

Anil_Victoria_Solution_Intern_ | Tableau Public

Visualization Observation: Scatter Plot - Relationship Between Discount Percentage and Revenue

Observation:

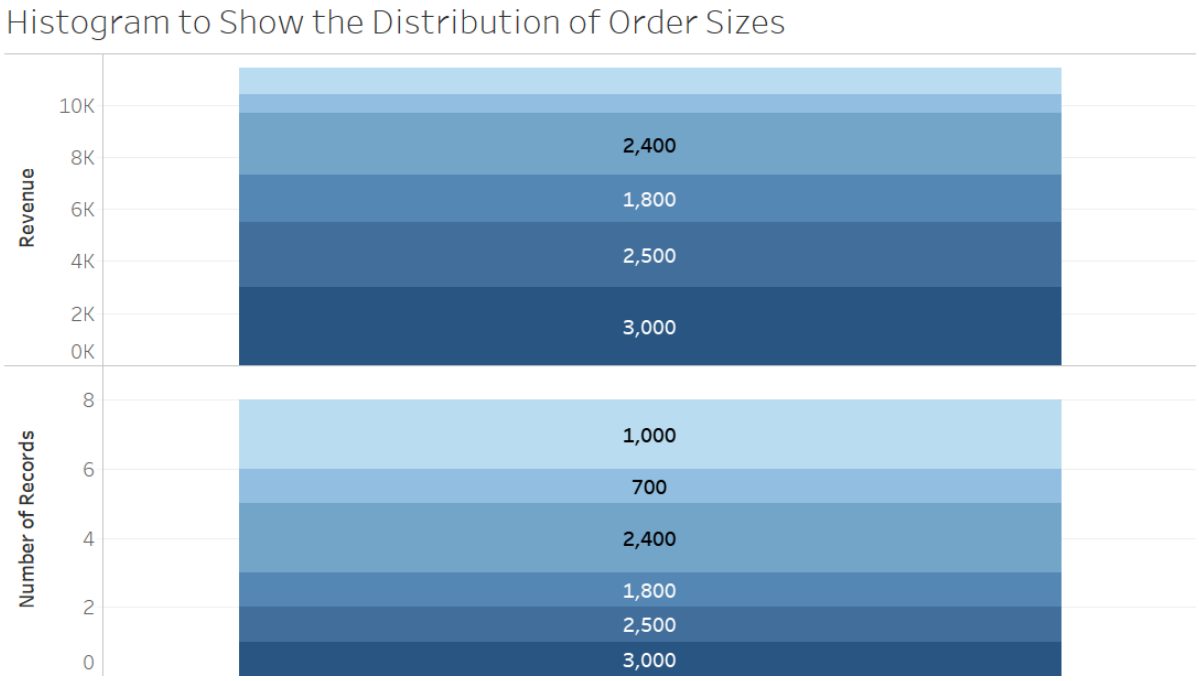
The scatter plot titled provides insights into the relationship between the discount percentage offered and the corresponding revenue generated. Key observations include:

1. **Higher Discounts and Higher Revenue**:
    - Data points with higher discounts (40%) are associated with higher revenues (£4,300 and £5,400).
2. **Lower Discounts and Lower Revenue**:
    - A data point with a lower discount (5%) is associated with a lower revenue (£1,700).

Tableau public Link - [Anil_Victoria_Solution_Intern__ | Tableau Public]

Business Insight:

- The plot suggests a positive correlation between discount percentage and revenue, indicating that offering higher discounts might lead to increased revenue. This insight could be valuable for developing discount strategies to optimize sales.



Observation:

The histogram visualizes the distribution of revenue and the number of records across different ranges. Key observations include:

Revenue Distribution:

- The revenue values are categorized into different ranges: £0K, £2K, £4K, £6K, £8K, and £10K.
- Notable revenue peaks include £3,000, £2,500, £1,800, and £2,400.

Number of Records:

- The number of records is categorized into different ranges: 0, 2, 4, 6, and 8.
- Significant revenue amounts include £3,000, £2,500, £1,800, £2,400, £700, and £1,000.

This visualization provides a clear comparison between the revenue generated and the number of records for different order sizes, helping to understand the distribution and frequency of order sizes in terms of both revenue and records.

Tableau public Link - https://public.tableau.com/app/profile/anil.kumbhar/viz/Anil_Victoria_Solution_Intern/HistogramtoShowtheDistributionofOrderSizes

Problem 4. **Final Deliverables Submit your work in the project cover sheet**

1. Cleaned dataset in Google Sheets or Excel.
2. SQL queries used for data cleaning and aggregation.
3. Three visualizations (Heatmap, Scatter Plot, Histogram).
4. Summary report detailing key findings and business recommendations.

Sol - **Final Deliverables Overview**

1. Cleaned Dataset:

- The dataset was cleaned by removing duplicates, handling missing values, and ensuring data consistency across columns such as **Revenue**, **Discount (%),** and **Product Category**.

2. **SQL queries used for data cleaning and aggregation.**

- **Total Revenue per Product Category:**
- **Furniture** and **Electronics** are the most profitable product categories, contributing significantly to the overall revenue. **Clothing** has a much lower total revenue, which could indicate that it is a less popular or lower-margin product segment.
- **Average Discount Applied Across Different Product Categories:**
- The **average discount** applied across product categories shows that **Furniture** had the highest average discount, indicating a strategy to drive sales for this category. On the other hand, **Clothing** had the lowest average discount, suggesting a less aggressive discounting strategy or potential pricing challenges.
- **Monthly Sales Trends:**
- **Sales performance** shows significant peaks in **January** and **February**, which could be attributed to seasonal demand or specific promotional events during these months. The relatively lower sales in other months may suggest potential opportunities for targeted marketing or promotions during off-peak months.
- **Distribution of Order Sizes:**
- The **distribution of order sizes** indicates that most orders are of smaller value, with fewer large orders. This suggests that there may be an opportunity to increase the number of large orders through strategies like **bundling products** or offering incentives for larger purchases.

3. Visualizations:

- **Heatmap**: Visualized revenue performance across different months and product categories, identifying profitable segments and seasonal sales patterns.

- **Scatter Plot**: Examined the relationship between **Discount (%)** and **Revenue**, showing how discounts influence sales performance across product categories.
- **Histogram**: Showed the distribution of order sizes, helping to identify common order sizes and pinpoint opportunities to boost large orders.

Key Findings:

1. **Total Revenue per Product Category**:
    - **Furniture** and **Electronics** are the most profitable product categories, contributing significantly to total revenue. These categories should be prioritized for future marketing and sales efforts.
    - **Clothing** generates the lowest revenue, indicating a need for strategic adjustments, such as promotions or improved product offerings.
2. **Average Discount Applied Across Product Categories**:
    - **Furniture** has the highest average discount of **20%**, followed by **Electronics** at **15%** and **Clothing** at **1.67%**.
    - The larger discounts on Furniture and Electronics may be contributing to higher sales, but it's important to monitor the impact on profit margins.
3. **Monthly Sales Trends**:
    - **January** and **February** have high sales, with January reaching **3500 units**, indicating a potential seasonal peak.
    - Sales dip in **March** and **April**, suggesting the need for promotional efforts or campaigns to sustain sales during slower months.
4. **Order Size Distribution**:
    - The majority of orders are clustered around smaller sizes, with fewer large orders. This may indicate that while there are many smaller purchases, opportunities exist to target customers for larger, bundled purchases.

Business Recommendations:

1. **Focus on High-Performing Categories**:
    - Continue to invest in **Furniture** and **Electronics** as these categories are the most profitable.
    - Offer targeted promotions and improve product offerings in **Clothing** to increase its sales contribution.
2. **Optimize Discount Strategy**:
    - Evaluate the profitability of discounts, especially for **Furniture** and **Electronics**. While these products benefit from larger discounts, consider offering **tiered discounts** for high-value customers to maintain margins.
    - Reduce discount percentage on lower-performing categories like **Clothing** while experimenting with other promotional strategies like bundling.

3. **Leverage Seasonal Trends**:
    - Capitalize on the sales peak in **January** and **February** by launching new products or running promotional campaigns during these months.
    - Implement strategies such as targeted ads or seasonal promotions to boost sales in **March** and **April**, when sales tend to decline.
4. **Increase Large Orders**:
    - Identify customers with the potential for larger orders by offering volume discounts or loyalty programs.
    - Use upselling and cross-selling techniques to encourage customers to purchase larger quantities or more expensive products.

This report provides actionable insights that can guide future business decisions for maximizing revenue and optimizing sales strategies across different product categories.