****

**COLLEGE OF COMPUTING AND INFORMATION SCIENCES**

**PROGRAMME**: INFORMATION SYSTEMS AND TECHNOLOGY

**COURSE**: SYSTEMS INTEGRATION AND DEPLOYMENT

XCOMPANY DOCUMENTATION

GROUP MEMBERS

|  |  |
| --- | --- |
| **Name** | **Reg No.** |
| Sseremba Julius | 21/U/09154/Eve |
| Wakayinja Frank | 21/U/07336/Eve |
| Opungo Jacob | 21/U/14200/Eve |
| Anzoekoku John Paul | 21/U/08691/Eve |
| Jjemba Lawrence | 21/U/19199/Eve |

**Data Merge and Integration Process Documentation**

**Objective:**

The objective of this data merge and integration process is to combine sales data with customer demographic information to conduct a comprehensive analysis. By merging these datasets, we aim to explore the relationship between sales figures and customer demographics such as age and country.

**Tools Used:**

* RStudio
* ggplot2
* scatterplot3d
* dplyr
* plotrix

**Datasets Involved:**

* **sales**: Contains sales data including ID and Sales columns.
* **customers**: Contains customer data including ID, Age, and Country columns.

**Steps Taken:**

Below are some of the steps take in the

**1. Data Collection:**

* The **sales** and **customers** datasets were loaded into RStudio for analysis.

2. Data Cleaning:

The library(dplyr) was used for cleaning the sales and customers datasets.

# Read the dirty sales data from an external CSV file

sales <- read.csv("C:/Users/HP/Desktop/x-company/x-company/unclean data/sales.csv")

# Read the dirty customers data from an external CSV file

customers <- read.csv("C:/Users/HP/Desktop/x-company/x-company/unclean data/customers.csv")

# Check the structure of the data frames

str(sales)

str(customers)

# Assume sales and customers are data frames containing your sales and customer data, respectively

# Check for missing values in sales table

missing\_sales <- sum(is.na(sales))

# Remove rows with missing values

clean\_sales <- na.omit(sales)

# Remove duplicates in sales table

clean\_sales <- distinct(clean\_sales)

# Standardize formats if needed (e.g., converting dates to a consistent format)

# Check for missing values in customers table

missing\_customers <- sum(is.na(customers))

# Remove rows with missing values

clean\_customers <- na.omit(customers)

# Remove duplicates in customers table

clean\_customers <- distinct(clean\_customers)

# Standardize formats if needed

# Specify the storage location for the cleaned files

clean\_sales\_file <- "C:/Users/HP/Desktop/x-company/x-company/clean\_sales.csv"

clean\_customers\_file <- "C:/Users/HP/Desktop/x-company/x-company/clean\_customers.csv"

# Save cleaned data to new files

write.csv(clean\_sales, file = clean\_sales\_file, row.names = FALSE)

write.csv(clean\_customers, file = clean\_customers\_file, row.names = FALSE)

# Optionally, you can combine sales and customers data after cleaning if needed

# cleaned\_data <- merge(clean\_sales, clean\_customers, by = "ID")

# Summary of cleaning process

cat("Cleaning Summary:\n")

cat("Sales Table:\n")

cat("Initial rows:", nrow(sales), "\n")

cat("Rows after removing missing values:", nrow(clean\_sales), "\n")

cat("Rows after removing duplicates:", nrow(distinct(clean\_sales)), "\n")

cat("\n")

cat("Customers Table:\n")

cat("Initial rows:", nrow(customers), "\n")

cat("Rows after removing missing values:", nrow(clean\_customers), "\n")

cat("Rows after removing duplicates:", nrow(distinct(clean\_customers)), "\n")

**2. Merge Strategy:**

* The **merge()** function was used to merge the **sales** and **customers** datasets based on the 'ID' column, combining relevant information from both datasets into a single dataframe called **mergedData**.
* The merge function;

mergedData<-merge(sales[, c("ID", "Sales")],customers[, c("ID", "Age", "Country")])

mergedData$Age\_Group <- cut(mergedData$Age, breaks = c(0, 20, 30, 40, 50, 60, Inf),

labels = c("0-20", "21-30", "31-40", "41-50", "51-60", "61+"))

**3. Data Exploration:**

* Summary statistics and a preview of the merged dataset were generated using functions such as **summary()** and **head()** to gain insights into the structure and content of the data.
* The functions used;

summary(mergedData)

head(mergedData)

**4. Data Visualization:**

* The **ggplot2** package was utilized to create visualizations of the merged dataset. Specifically, a bar plot was generated to visualize sales figures by country, providing a graphical representation of the relationship between sales and customer geography.
* The functions used;

ggplot(mergedData, aes(x = Age)) +

geom\_histogram(binwidth = 5, fill = "skyblue", color = "black") +

labs(title = "Customer Age Distribution", x = "Age", y = "Frequency") +

theme\_minimal()

ggplot(mergedData, aes(x = Age\_Group, y = Sales)) +

geom\_bar(stat = "summary", fun = "sum", fill = "lightgreen") +

labs(title = "Sales by Age Group", x = "Age Group", y = "Total Sales") +

theme\_bw()

ggplot(mergedData, aes(x = Country, y = Sales, fill = Age\_Group)) +

geom\_bar(stat = "summary", fun = "sum", position = "dodge") +

labs(title = "Sales by Country and Age Group", x = "Country", y = "Total Sales", fill = "Age Group") +

theme\_minimal()

ggplot(mergedData, aes(x = Country)) +

geom\_bar(fill = "salmon") +

labs(title = "Customer Distribution by Country", x = "Country", y = "Number of Customers") +

theme(axis.text.x = element\_text(angle = 45, hjust = 1)) # Rotate x-axis labels for better readability

**5. Session Information:**

* The **sessionInfo()** function was used to capture information about the R environment and package versions used during the analysis, ensuring reproducibility and transparency.
* The information about the R session;

* loaded via a namespace (and not attached):
* [1] vctrs\_0.6.5 cli\_3.6.2 rlang\_1.1.3 generics\_0.1.3 glue\_1.7.0 labeling\_0.4.3
* [7] colorspace\_2.1-0 scales\_1.3.0 fansi\_1.0.6 grid\_4.3.3 munsell\_0.5.1 tibble\_3.2.1
* [13] lifecycle\_1.0.4 compiler\_4.3.3 dplyr\_1.1.4 pkgconfig\_2.0.3 farver\_2.1.1 lattice\_0.22-5
* [19] R6\_2.5.1 tidyselect\_1.2.1 utf8\_1.2.4 pillar\_1.9.0 magrittr\_2.0.3 tools\_4.3.3
* [25] withr\_3.0.0 gtable\_0.3.5

**Decisions Made:**

The decision to merge the datasets using the 'ID' column as the key variable was deliberate. By linking sales records with corresponding customer demographics through a unique identifier, we aimed to facilitate a comprehensive analysis of the relationship between sales performance and customer characteristics. This approach allows us to precisely associate each sale with its respective customer, enabling deeper insights into purchasing behavior and customer segmentation. Additionally, utilizing the 'ID' column as the key variable ensures accuracy and consistency in the merged dataset, laying a solid foundation for meaningful analysis and decision-making processes.

**Challenges Encountered:**

1. **Identifying R Packages:**
   * One challenge involved identifying and selecting the appropriate R packages for data manipulation, visualization, and analysis. This required research and experimentation to find the most suitable packages for the task at hand.
2. **Learning R Syntax:**
   * Another challenge was becoming proficient in R syntax, particularly for data manipulation tasks such as merging datasets and creating visualizations. Overcoming this challenge involved investing time in learning R programming concepts and syntax through online resources and tutorials.
3. **Obtaining Datasets:**
   * Acquiring the necessary datasets posed a challenge, as it required locating reliable sources and obtaining permission to use the data. Additionally, ensuring that the datasets were in a format compatible with R and free from errors or inconsistencies added complexity to the process.
4. **Cleaning and Preprocessing Datasets:**
   * Cleaning and preprocessing the datasets presented challenges, including handling missing values, addressing data inconsistencies, and ensuring data quality. This required careful attention to detail and the implementation of appropriate data cleaning techniques to prepare the datasets for merging.
5. **Merging Datasets:**
   * Merging the datasets proved to be challenging due to differences in data formats, variable names, and unique identifiers. Ensuring that the merge operation was conducted accurately and efficiently required a thorough understanding of the data structures and merge functions available in R.

**Results:**

* The merged dataset, **mergedData**, combines sales data with customer demographic information, enabling further analysis and exploration of the relationship between sales figures and customer demographics.