R

Cleaning

- The numeric value of the months has been changed to text

```
7df <- df %>%
+ mutate(Month_name = month.abb[Month])
```

- The numeric value of the hours has been changed to text

```
df <- df %>%
mutate(Day_of_week_name = case_when(
year_frequency <- df %>%
filter(Transaction_type == "Sale") %>%
group_by(Year, Month_name) %>%
summarise(monthly_sales = sum(Total), .groups = "drop")
Day_of_week == 1 ~ "Sunday",
Day_of_week == 2 ~ "Monday",
Day_of_week == 3 ~ "Tuesday",
Day_of_week == 4 ~ "Wednesday",
Day_of_week == 5 ~ "Thursday",
Day_of_week == 6 ~ "Friday",
Day_of_week == 7 ~ "Saturday",
TRUE ~ NA_character_
))
```

- Create a time slot

```
- df <- df %>%
mutate(timeslot = case_when(
Hour >= 6 & Hour < 12 ~ "Morning",
Hour >= 12 & Hour < 18 ~ "Evening",
Hour >= 18 & Hour < 24 ~ "Night",
TRUE ~ "Early morning"
))
percentage_timeslot <- df %>%
group_by(timeslot) %>%
summarise(ventas = sum(Total, na.rm = TRUE)) %>%
mutate(porcentaje = (ventas / sum(ventas)) * 100)
```

Analyse

- General dataset exploration

```
str(df)
summary(df)
skimr::skim(df) # skimr para resumen más complete
```

- Analysis: Top 10 products by quantity sold (excluding returns)

- Formula:

```
top10_products <- df %>%
group_by(Description) %>%
summarise(Total_Quantity = sum(Quantity)) %>%
arrange(desc(Total_Quantity)) %>%
head(10)
```

-Analysis: Top Customers

-Formula:

```
top_10customers <- df %>%
+ filter(Transaction_type == "Sale") %>%
+ group_by(customerID_) %>%
+ summarise(revenue = sum(Total), .groups = "drop") %>%
+ arrange(desc(revenue)) %>%
+ head(10)
sum_sales_topcustomers <- sum(top_10customers$revenue)
sum_sales_total <- sum(df$Total)
9percentage_top_customers <- (sum_sales_topcustomers/ sum_sales_total) * 100
Dataframe= top10.customerscountry
top_10_info <- top_10customers %>%
inner_join(df %>% select(customerID_, Country) %>% distinct(), by = "customerID_")
```

-Analysis: Which countries generate the most money?
-Formulas:

```
top_countries <- df %>%
filter(Transaction_type == "Sale") %>%
group_by(Country) %>%
summarise(total_revenue = sum(Total), .groups = "drop") %>%
arrange(desc(total_revenue))
dataframe = top10_info
dataframe= customers_per_country; top_countries
```

- Analysis: Months with the most sales
-Formula:

```
sales_per_day <- df %>%
filter(Transaction_type == "Sale") %>%
group_by(Day_of_week_name) %>%
summarise(total_sales = sum(Total), .groups = "drop") %>%
arrange(desc(total_sales))
```

Fórmula para saber que representa del total las vents de un día:

```
> sum(df$Total)
[1] 8280356
thursday <- (1971822/ 8280356) * 100
> tuesday <- (1697055.6/ 8280356) * 100
```

Dataframes: sales_per_day
Datafrma: sales_per_hour, percentage time

-Analysis: What percentage of transactions are
returns? And Return Country
-Formulas:

```
total_sales <- nrow(filter(df, Transaction_type == "Sale"))
total_returns <- nrow(filter(df, Transaction_type == "Return"))
return_rate <- total_returns / (total_sales + total_returns)
percentage_return <- (340 / sum(df$Transaction_type == "Return")) * 100
> View(percentage_return)
percentage_return_country <- (7476/ sum(df$Transaction_type == "Return" )) * 100
return_rate_country <- df %>%
group_by(Country) %>%
summarise(
12Sales = sum(Transaction_type == "Sale"),
Return = sum(Transaction_type == "Return"),
Return_Rate = Return / (Sales + Return),
.groups = "drop"
) %>%
arrange(desc(Return_Rate))
return_rate_product <- df %>%
group_by(Description) %>%
summarise(
Sales = sum(Transaction_type == "Sale"),
Return = sum(Transaction_type == "Return"),
Return_Rate = Return / (Sales + Return),
.groups = "drop"
) %>%
arrange(desc(Return_Rate))
```

-Analysis: Customer segmentation
-Formulas:

```
# Calcular RFM por cliente
rfm <- df %>%
filter(Transaction_type == "Sale") %>%
group_by(customerID_) %>%
summarise(
Recency = as.numeric(difftime(analysis_date, max(InvoiceDate), units = "days")),
Frequency = n_distinct(InvoiceNo),
Monetary = sum(Total),
.groups = "drop"
)
# Para segmentar, puedes crear cuartiles o quintiles, por ejemplo:
rfm <- rfm %>%
mutate(
R_Score = ntile(-Recency, 4), # Más reciente es mejor, por eso negativo
F_Score = ntile(Frequency, 4),
```

```
M_Score = ntile(Monetary, 4),
RFM_Score = R_Score + F_Score + M_Score
)

top_customers <- rfm %>% arrange(desc(RFM_Score))
top_rfm_count <- sum(rfm$RFM_Score == 12)
total_clients <- nrow(rfm)
percentage_top_rfm <- (top_rfm_count / total_clients) * 100
valuable_customers <- rfm_info %>%
filter(RFM_Score == 12) %>%
group_by(Country) %>%
summarise(Valuable_customers = n()) %>%
arrange(desc(Valuable_customers))
rfm_info %>%
mutate(Grupo = ifelse(RFM_Score == 12, "Top", "Resto")) %>%
group_by(Grupo) %>%
summarise(Revenue = sum(Monetary))
top_rfm_count <- sum(rfm$RFM_Score == 12)
total_clients <- nrow(rfm)
percentage_top_rfm <- (top_rfm_count / total_clients) * 100
valuable_customers <- rfm_info %>%
filter(RFM_Score == 12) %>%
group_by(Country) %>%
summarise(Valuable_customers = n()) %>%
arrange(desc(Valuable_customers))
rfm_info %>%
mutate(Group = ifelse(RFM_Score == 12, "Top", "Others")) %>%
group_by(Group) %>%
summarise(Revenue = sum(Monetary))
rfm_info_category <- rfm_info %>%
mutate(Segmento = case_when(
RFM_Score >= 10 ~ "Excellent Customers",
RFM_Score >= 7 ~ "Loyal Customers",
RFM_Score >= 4 ~ "Risk Customers",
TRUE ~ "Lost Customers"
```

- Analysis: Segmentation by country

-Formula:

```
customers_per_country <- df %>%
filter(Transaction_type == "Sale") %>%
group_by(Country) %>%
summarise(
Total_Customers = n_distinct(customerID_),
Total_Sales = sum(Total),
.groups = "drop"
) %>%
arrange(desc(Total_Sales))
```

R graphics

```
df %>%
group_by(Year, Month) %>%
summarise(monthly_sales = sum(Total), .groups = "drop") %>%
ggplot(aes(x = interaction(Year, Month, sep = "-"), y = monthly_sales)) +
geom_line(group = 1, color = "steelblue") +
labs(title = "Sales per Month",
x = "Month",
y = "Total Sales (£)") +
theme_minimal() +
theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

- Key Performance Indicators (KPIs)

The company generated over £8.28M in total sales, with 4,371 unique customers during the analyzed period.

Out of these, 474 customers scored the highest possible RFM value (12/12) — signaling key contributors to revenue.

The overall return rate was 2.2%, with returns primarily concentrated in a few

specific countries and products.
These metrics offer a quick, high-level overview of business performance and will be referenced throughout the analysis.