# E-commerce Furniture Dataset 2024

- **Tools used**- SQL, R Studio

- **Objective-** The objective of this project is to analyze sales patterns in an e-commerce furniture dataset using SQL and R. SQL is used for data cleaning and transformation, while R is employed for visualizations. The goal is to identify key factors that influence sales—such as price, discounts, and shipping tags.

- **Dataset Overview-** Contains ~2,000 furniture items.

  Key columns

- productTitle
- originalPrice (lots of missing values)
- price (current selling price)
- sold (target variable)
- tagText (categorical — like "Free shipping", etc.)

- **Initial Data cleaning has been done in SQL**

**SQL Code-**

```
            use FURNITURE;

SELECT
    productTitle,

    -- Sanitize and convert originalPrice
    TRY_CAST(REPLACE(REPLACE(LTRIM(RTRIM(originalPrice)), '$', ''), ',', '') AS FLOAT) AS
originalPrice,

    -- Sanitize and convert price
    TRY_CAST(REPLACE(REPLACE(LTRIM(RTRIM(price)), '$', ''), ',', '') AS FLOAT) AS price,

    sold,

    -- Simplify tagText categories
    CASE
        WHEN tagText = 'Free shipping' THEN 'Free shipping'
        WHEN tagText = '+Shipping: $5.09' THEN '+Shipping: $5.09'
        ELSE 'others'
    END AS tagText,

    -- Calculate discount percentage with null-safe conversion
    ROUND(
        ((TRY_CAST(REPLACE(REPLACE(LTRIM(RTRIM(originalPrice)), '$', ''), ',', '') AS FLOAT) -
          TRY_CAST(REPLACE(REPLACE(LTRIM(RTRIM(price)), '$', ''), ',', '') AS FLOAT)) /
          NULLIF(TRY_CAST(REPLACE(REPLACE(LTRIM(RTRIM(originalPrice)), '$', ''), ',', '') AS FLOAT),
0)) * 100.0, 2
    ) AS discount_percentage

INTO FURNITURE.dbo.furniture_cleaned
FROM FURNITURE.dbo.ecommerce_furniture_dataset_2024
WHERE originalPrice IS NOT NULL
  AND ISNUMERIC(REPLACE(REPLACE(LTRIM(RTRIM(originalPrice)), '$', ''), ',', '')) = 1
  AND ISNUMERIC(REPLACE(REPLACE(LTRIM(RTRIM(price)), '$', ''), ',', '')) = 1;
---
SELECT
    tagText,
    COUNT(*) AS item_count,
    ROUND(AVG(discount_percentage), 2) AS avg_discount
```

```
FROM FURNITURE.dbo.furniture_cleaned
GROUP BY tagText
ORDER BY avg_discount DESC;


SELECT * FROM FURNITURE.dbo.furniture_cleaned;
```

| Shipping Tag | Product Count | Avg Discount (%) |
|---|---|---|
| Free shipping | 485 | 47.24% |
| others | 2 | 40.41% |

- ***Free shipping items dominate*** *the dataset and offer a* ***higher average discount*** *than others.*

  - **R has been used for advanced analysis and visualisation**

**R Code-**

```
# Load necessary packages

library(tidyverse)

library(ggplot2)

library(readr)

library(caret)

# Load cleaned dataset

furniture = read_csv("C:\\Users\\intel\\OneDrive\\Desktop\\unified mentor\\E-commerce furniture Dataset\\Furniture_Cleaned.csv")

furniture <- read_csv("C:\\Users\\intel\\OneDrive\\Desktop\\unified mentor\\E-commerce furniture Dataset\\Furniture_Cleaned.csv", col_names = FALSE)

# Manually set the column names

colnames(furniture) <- c("productTitle", "originalPrice", "price", "sold", "tagText", "discount_percentage")

# View structure

glimpse(furniture)

##EDA

# Summary statistics

summary(furniture)

# Check for NA values

colSums(is.na(furniture))

# Distribution of sold values

ggplot(furniture, aes(x = sold)) +
```

```
  geom_histogram(fill = "#2c3e50", bins = 30, color = "white") +
  labs(title = "Distribution of Furniture Items Sold", x = "Units Sold", y = "Count")
```

**##Analyse discount impact**

**# Scatter plot: Discount vs. Sold**

```
ggplot(furniture, aes(x = discount_percentage, y = sold)) +
  geom_point(color = "#e74c3c", alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Discount % vs. Items Sold", x = "Discount Percentage", y = "Units Sold")
```

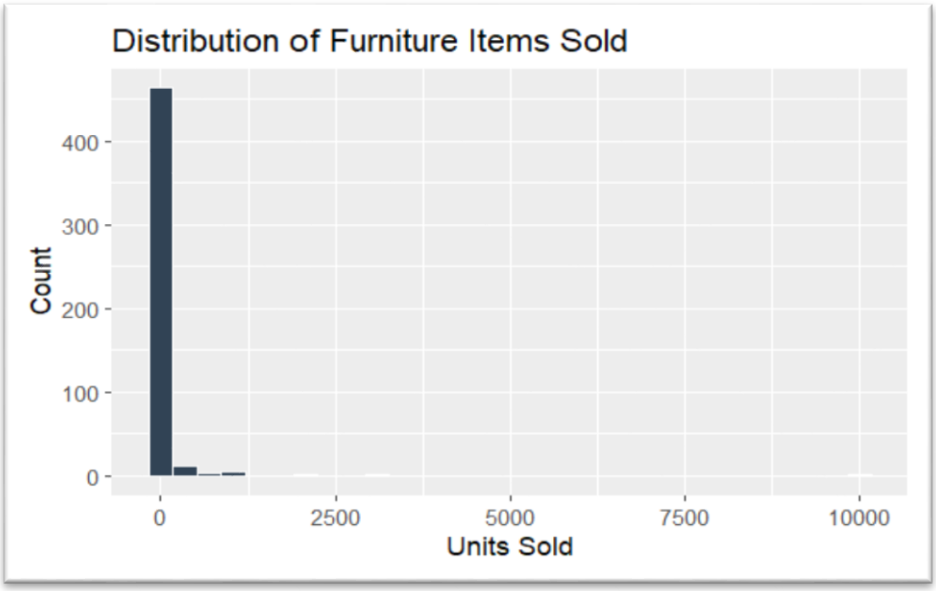
**# Boxplot: Shipping Type vs. Sold**

```
ggplot(furniture, aes(x = tagText, y = sold, fill = tagText)) +
  geom_boxplot() +
  labs(title = "Items Sold by Shipping Tag", x = "Shipping Tag", y = "Units Sold") +
  theme_minimal()
```

**# Correlation matrix for numeric features**

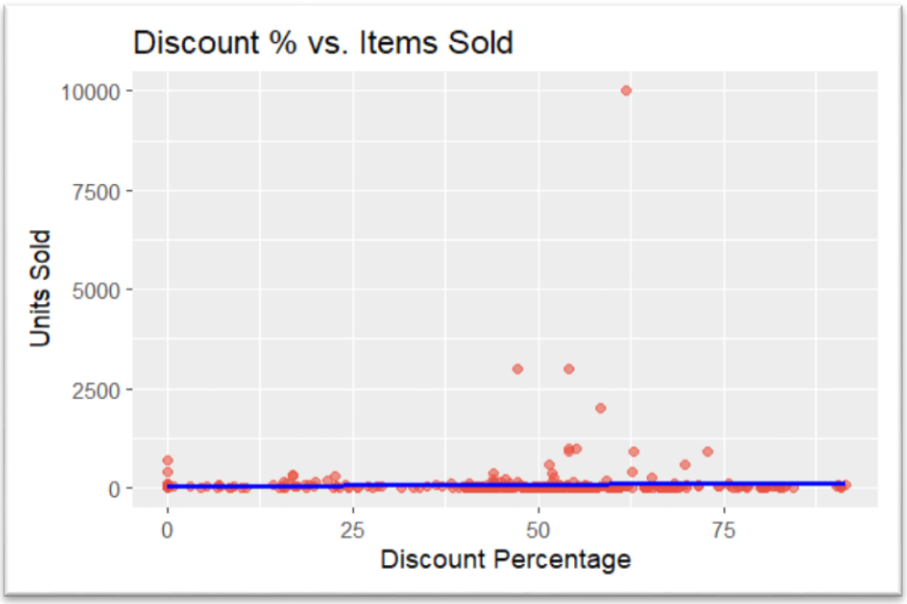
```
numeric_data <- furniture %>% select(price, originalPrice, discount_percentage, sold)
cor_matrix <- cor(numeric_data, use = "complete.obs")
corrplot(cor_matrix, method = "circle", type = "lower", tl.cex = 0.8)
```

- **Output plots-**

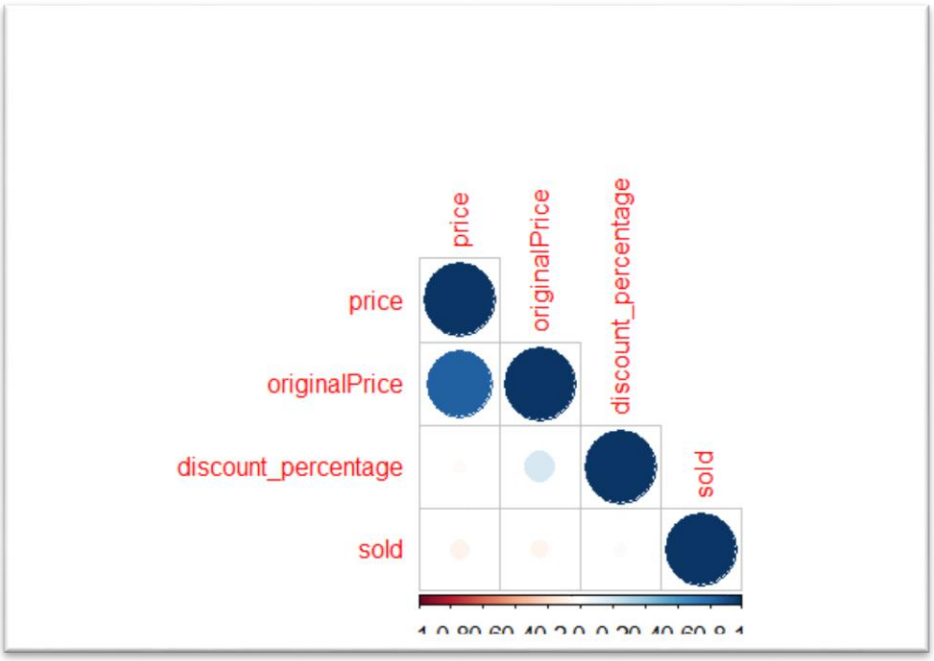



- *Distribution of units sold showing a high frequency of low-sales items with few outliers having large sales.*


- *Boxplot comparing items sold across different shipping tag categories, showing higher sales variability within 'Free shipping'.*




- *Visualizing the relationship between discount percentage and items sold, suggesting a slight upward trend with significant variability.*




- *Correlation heatmap revealing weak correlations between numeric features, indicating limited linear relationships.*

### ➢ Conclusion

This project successfully explored and analyzed an e-commerce furniture dataset using SQL for preprocessing and R for detailed exploratory analysis. Key findings revealed that:

- The majority of products offer "Free shipping" and are associated with higher average discounts.
- Discount percentage appears to influence sales to some extent, though other hidden factors likely play significant roles.
- Shipping tags also show variations in units sold, suggesting promotional strategies could be optimized.
- Correlation between numeric features (price, originalPrice, discount, sold) was relatively weak, reinforcing the idea that other non-numeric or contextual factors may be at play.