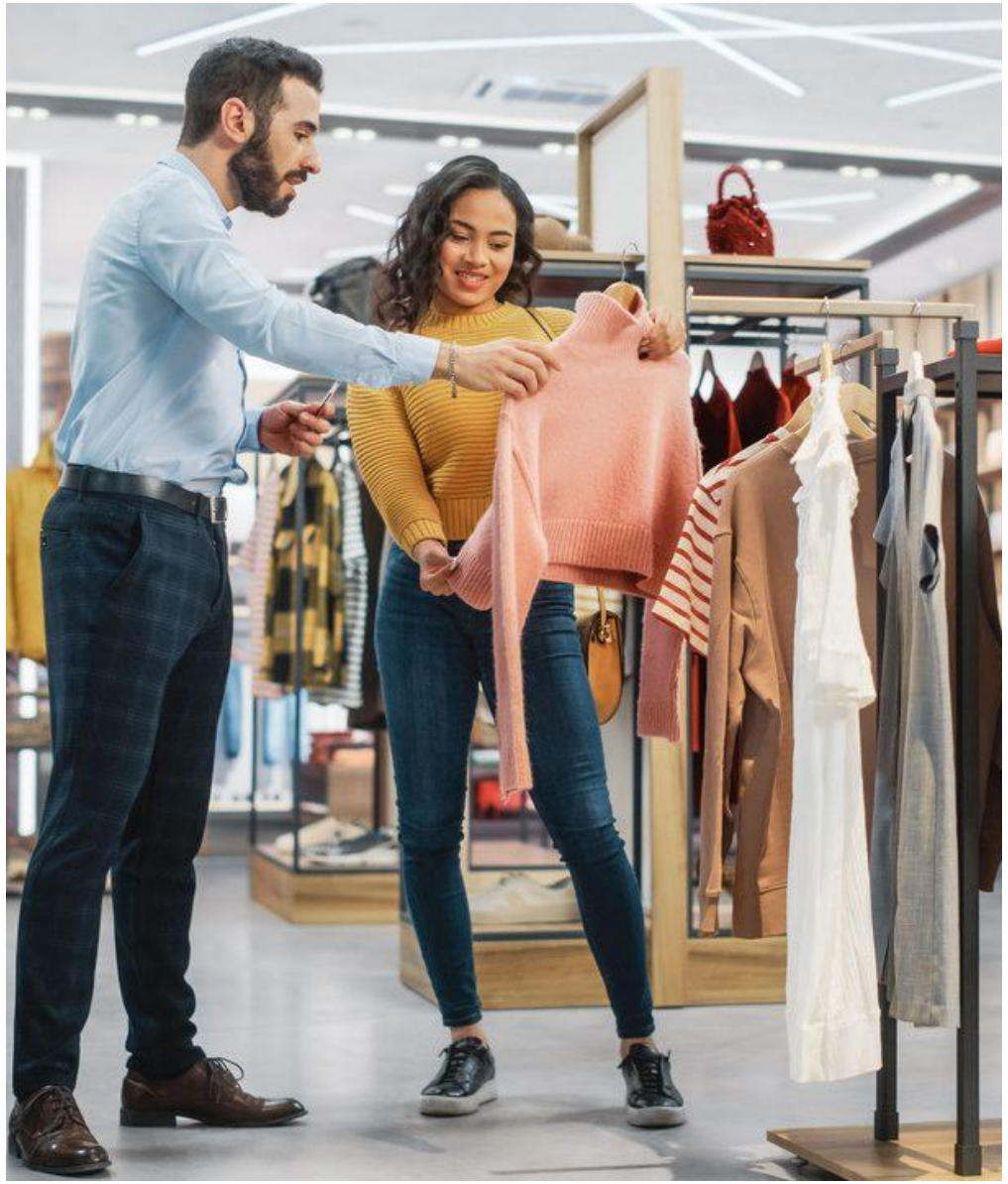
Project Report - Retailing Study

Duong Thuy Le



Part 1: Data Management

- MySQL
- Focusing on Fashion Retailing

Business Understanding

Business Application Area

Sales and Revenue Analysis for a Clothing Company:

We have chosen an online retail store's product purchase transactions for a clothing company as the business application area for developing the database.

The purpose of the database:

To manage various aspects of the retail store's operations, including:

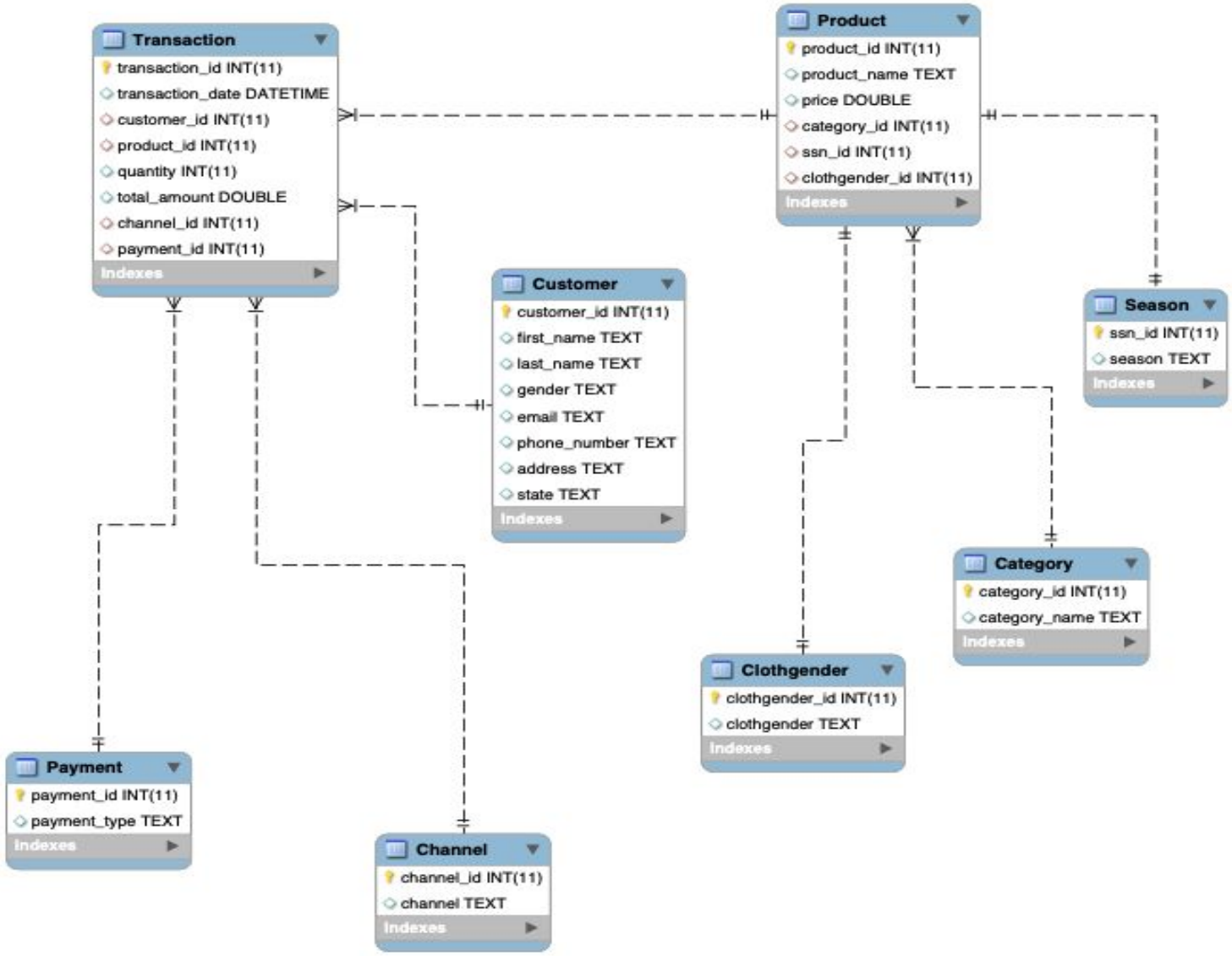
Data Preparation and Understanding

Data Understanding - Tables

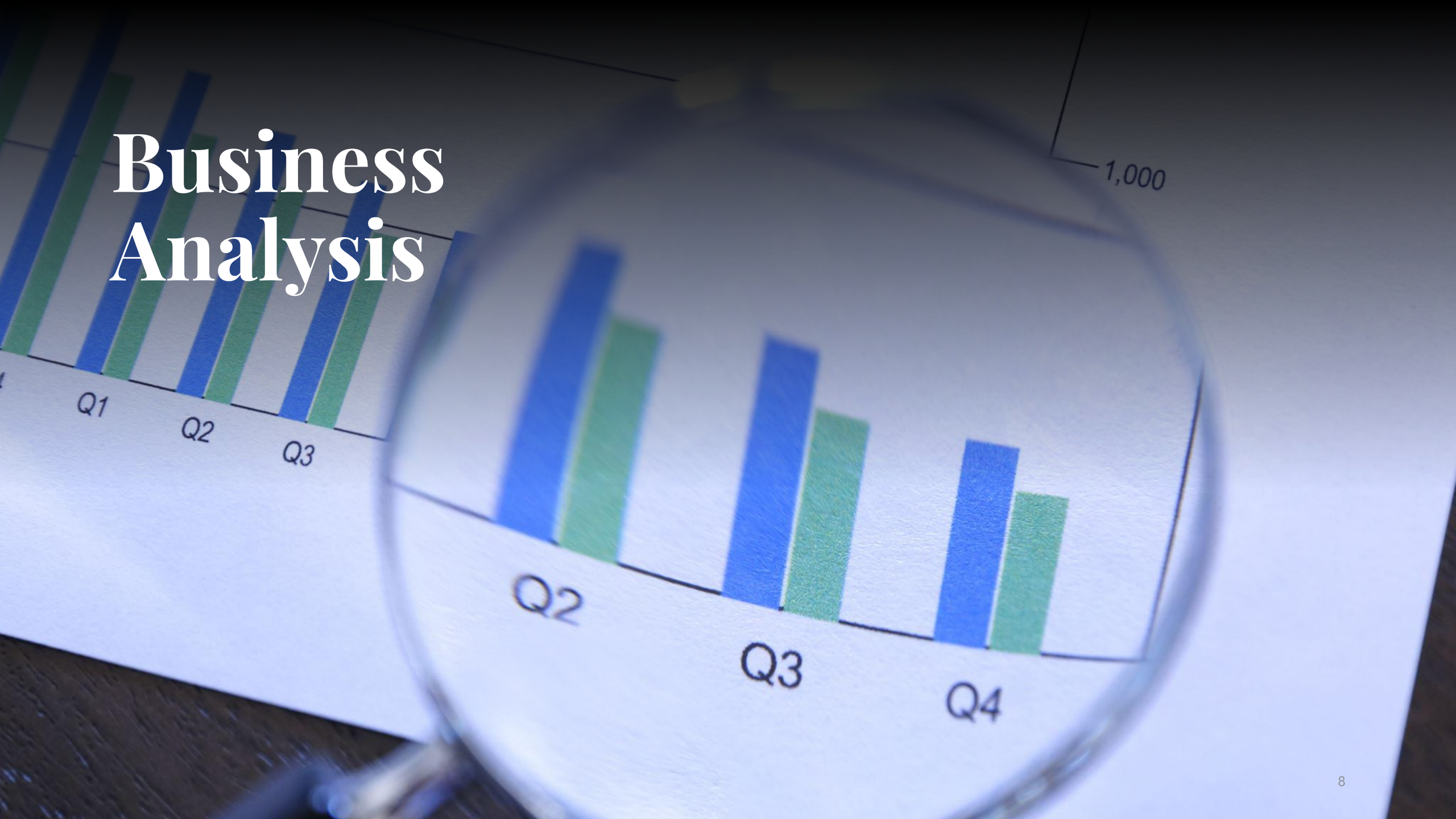
There are 8 tables in the database:

1. **"Transaction" table** stores information related to customer purchases, such as transaction IDs, customer IDs, product IDs, quantities, prices, and payment types.
2. **"Customer" table** stores customer information, such as customer IDs, names, gender, addresses, contact information, and purchase history.
3. **"Products" table** stores information about the available clothing products, including product IDs, names, descriptions, prices, and cloth gender.
4. **"Season" table** stores seasonal promotions and discount information, such as season IDs and names.
5. **"Category" table** stores information about product categories, such as category IDs and names.
6. **"Clothgender" table** would store information about gender-specific clothing items, such as clothgender IDs and descriptions.
7. **"Channel" table** would store information about sales channels, such as channel IDs and descriptions.
8. **"Payment" table** would store information related to payment processing, such as payment IDs, transaction IDs, and payment methods.

Data Understanding - EER Diagram



Business Analysis



Business Questions

SQL Queries will combine 11 business aspects:

Seasonal Business Performance Evaluation

Q1) What is the total revenue generated by each season?

Analysis:

It can be observed from the results that among all seasons, “Fall” generates the highest revenue and “Spring” generates the lowest revenue respectively. This indicates that seasonality has a significant impact on customer behavior, in order to increase revenue generation per season business should make informed decisions considering various factors i.e. adapting to operational efficiency in the busiest time and putting lucrative product offers in the slowest business period.

```
4   #1 What is the total revenue generated by each season?
5   SELECT Season.Season, ROUND(SUM(Transaction.total_amount),2) AS Revenue
6   FROM Transaction
7   JOIN Product ON Transaction.product_id = Product.product_id
8   JOIN Season ON Product.ssn_id = Season.ssn_id
9   GROUP BY Season.Season
10  ORDER BY Revenue DESC;
```

100% 55:4

Result Grid Filter Rows: Search Export:

Season	Revenue
Fa	83992.78
Wi	60624.57
Su	54100.51
Sp	3358.05

Payment Processing

Q2) What is the average payment amount by payment method?

Analysis:

It can be clearly seen from the output that “Cash” is the most commonly used and “Visa” is the least common payment method used by the customers. This could be inferred as despite the digital advancement, Cash is yet the most popular payment method, thus business should keep accepting the payments made by Cash while keeping digital payment methods available too. This can also help in improving customer experience.

```
12    #2.  What is the average payment amount by payment method?
13 •  SELECT Payment.payment_type AS Payment_type, ROUND(AVG(Transaction.total_amount),2) AS Average_payment
14    FROM Transaction
15    JOIN Payment ON Transaction.payment_id = Payment.payment_id
16    GROUP BY Payment.payment_type
17    ORDER BY Average_payment DESC;
```

100% 1:11

Result Grid Filter Rows: Search Export:

Payment_type	Average_payment
Cash	212.04
Credit Card	206.78
Coupon	198.74
Mastercard	197.91
VISA	196.38

Category Performance Evaluation

Q3) What is the total revenue generated by each category?

Analysis:

It can be illustrated from the output that “Category 1 - Outer” is generating the highest revenue and “Category 6 - Accessories” generates the lowest revenue respectively. Category 1 is the best performing category, business can evaluate and promote the product listing of it to increase more sales in this category. Similarly, Category 6 can be reviewed to optimize the product assortment, i.e. new product offerings can be added or the existing products that have the least sales contribution can be eliminated.

```
#3. What is the total revenue generated by each category?
SELECT Category.category_id AS Category_ID, Category.category_name AS Name, ROUND(SUM(Transaction.total_amount),2) AS Revenue
FROM Transaction
JOIN Product ON Transaction.product_id = Product.product_id
JOIN Category ON Product.category_id = Category.category_id
GROUP BY Category.category_id
ORDER BY Revenue DESC;
```

100% 58:19

Result Grid Filter Rows: Search Export:

Category_ID	Name	Revenue
00000000001	outer	68141.05
00000000004	cut & sewn	58313.97
00000000002	bottoms	42918.65
00000000003	shirts	12445.14
00000000007	inner & living	9629.05
00000000005	knit	9372.35
00000000009	dress	819.35
00000000006	accessories	436.35

Customer Segmentation

Q4) What is the total revenue generated by each gender?

Analysis:

It can be observed from the results that 60.3% of revenue is generated by “Men” and 35.6% by “Women” respectively. Business should offer a wide product assortment for men while tailor marketing campaigns that cater to men’s preference. Overall business planning should certainly take into account the gender role in revenue generation.

```
#4. What is the total revenue generated by each gender?
SELECT CG.clothgender AS Gender, ROUND(SUM(T.total_amount),2) AS Revenue
FROM Transaction T
JOIN Product P ON T.product_id = P.product_id
JOIN Clothgender CG ON P.clothgender_id = CG.clothgender_id
GROUP BY CG.clothgender
ORDER BY Revenue DESC;
```

100% 56:27

Result Grid Filter Rows: Search Export:

Gender	Revenue
Men	121939.94
Women	72062.12
Kid	7953.10
Others	120.75

Customer Loyalty

Q5) Which customers have made five or more transactions, and what is the total quantity and revenue generated from those transactions?

Analysis:

Loyal customers can be clearly seen from the results, now after identifying them personalized marketing strategies can be developed to target them. This would impact purchase frequency, increase retention rate and customer satisfaction. With the spread of positive word of mouth by loyal customers, brand image of the business would also be enhanced.

```
35  #5. Which customers have made five or more transactions, and what is the total quantity and revenue generated from those transactions?
36  SELECT c.first_name, c.last_name, COUNT(t.transaction_id) AS num_transactions,
37  SUM(t.quantity) AS total_quantity, ROUND(SUM(t.total_amount),2) AS total_revenue
38  FROM Transaction t
39  JOIN Customer c ON t.customer_id = c.customer_id
40  WHERE c.customer_id IN (
41    SELECT customer_id
42    FROM Transaction
43    GROUP BY customer_id
44    HAVING COUNT(*) >= 5
45  )
46  GROUP BY c.first_name, c.last_name
47  ORDER BY total_revenue DESC;
```

100% 21:47

Result Grid Filter Rows: Search Export:

first_name	last_name	num_transactions	total_quantity	total_revenue
Lacie	Shynn	5	29	1198.55
Jeanine	Heiner	5	29	813.69
Max	Niche	5	24	673.80
Sheryl	Arnaud	5	23	661.74
Tabby	Burkett	5	17	324.15

Customer Purchase Frequency

Q6) Which customers have made the most purchases in the last year?

Analysis:

Customers with a frequent purchase are highlighted in the output, business can cater to their specific needs to increase the basket size and develop strategies that lead to retaining them. This can help in improving revenue generation and overall business performance.

```
49    #6. Which customers have made the most purchases in the last year?
50 •  SELECT c.first_name, c.last_name, COUNT(t.transaction_id) AS num_transactions
51    FROM Transaction t
52    JOIN Customer c ON t.customer_id = c.customer_id
53    WHERE t.transaction_date >= DATE_SUB(NOW(), INTERVAL 1 YEAR)
54    GROUP BY c.first_name, c.last_name
55    ORDER BY num_transactions DESC
56    LIMIT 10;
```

100% 33:59

Result Grid | Filter Rows: Search | Export: | Fetch rows:

first_name	last_name	num_transactions
Rici	Foulis	4
Lacie	Shynn	4
Gusta	Fazan	3
Mile	Skells	3
Pepito	Stert	3
Sioux	Franklyn	3
Maryrose	Ghent	3
Agace	Whight	3
Tabby	Burkett	3
Marylou	McQuaid	3

Inactive Customers

Q7) Which customers have not made any purchases in the last six months?

Analysis:

Inactive customers are highlighted in the output; businesses can cater to their specific needs to increase the basket size and develop strategies that lead to retaining them. This can help improve revenue generation by retaining customers and overall business engagement.


```sql
# 7. Which customers have not made any purchases in the last six months?
SELECT c.first_name, c.last_name
FROM Customer c
LEFT JOIN Transaction t
ON c.customer_id = t.customer_id
WHERE t.transaction_date IS NULL OR t.transaction_date < DATE_SUB(NOW(), INTERVAL 6 MONTH)
GROUP BY c.customer_id
HAVING COUNT(t.transaction_id) = 0;
```

100% 36:65

Result Grid Filter Rows: Search Export:

first_name	last_name
Emmye	Shearman
Alyssa	Aisman
Vannie	Boddy
Moyra	Rowlinson
Florette	Flores
Giacobo	Mouse
Rona	Harnes
Ronalda	Rewcassell
Evania	Woodrough
Mandie	Hamshere
Rhianna	Moscone
Talbot	Petr
Symon	Yon
Malanie	Mazey
Erl	Armin
Jeremie	Reiners
Agustin	Scotchforth
Olivette	Credland
Raye	Littler
Nikola	Goskar
Francois	Cleeve
Konstanze	Duddy
Merrill	Gowthorpe
Norton	Waith
Cheslie	Boosey

Average Basket Value

Q8) What is the average purchase amount per customer?

Analysis:

The query helps us understand the average spending per customer per year. On average, customers spend $150 - $250 per year. This can extend to market basket analysis to understand purchase patterns, seasonality, popular items, etc.


```sql
# 8. What is the average purchase amount per customer?
SELECT c.first_name, c.last_name, ROUND(AVG(t.total_amount),2) AS avg_purchase_amount
FROM Transaction t
JOIN Customer c ON t.customer_id = c.customer_id
WHERE t.transaction_date >= DATE_SUB(NOW(), INTERVAL 1 YEAR)
GROUP BY c.first_name, c.last_name
ORDER BY avg_purchase_amount DESC;
```

100% 1:66

Result Grid Filter Rows: Search Export:

first_name	last_name	avg_purchase_amo...
Harvey	Brigman	1049.65
Ross	Salling	899.70
Roscoe	Depper	849.50
Tish	Hatrick	799.60
Rosemaria	Vales	799.60
Georgeanna	Stapylton	764.55
Marsha	Gaw	764.55
Wolfy	Simons	749.75
Claudine	Blankau	724.58
Ulises	Lamburn	599.80
Tyler	Bignell	599.80
Dolley	Peyton	599.60
Ardyth	Greatreax	594.65
Nomi	Dockrill	578.77
Emalia	Slyvester	524.65
Devlen	Josefsson	524.65
Elise	Coggen	524.65
Kamillah	Keirl	505.88
Sybyl	Garbutt	499.50
Cordi	Oag	499.50
Elna	Giraldon	499.50
Marylou	McQuaid	476.30
Harbert	Rhys	449.85
Doug	Nuzzetti	449.85
Darelle	Afonso	449.55

Best-Selling Products

Q9) What are the top 10 best-selling products by revenue and quantity sold, and how much revenue and quantity have they generated in total?

Analysis:

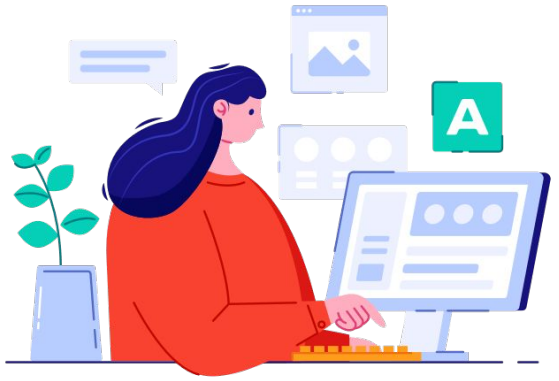
The top 10 selling products are highlighted in the output; The business can optimize inventory and improve the supply chain network, ensuring that these ten products are readily available to customers ensuring customer satisfaction. They can even strategize their marketing initiatives accordingly.

```
76    #9. What are the top 10 best-selling products by revenue and quantity sold, and how much revenue and quantity have they generated in total?
77 •  SELECT p.product_name, SUM(t.quantity) AS total_quantity, ROUND(SUM(t.total_amount),2) AS total_revenue
78    FROM Transaction t
79    JOIN Product p ON t.product_id = p.product_id
80    GROUP BY p.product_name
81    ORDER BY total_revenue DESC, total_quantity DESC
82    LIMIT 10;
```

100% 1:75

Result Grid | Filter Rows: Search | Export: | Fetch rows:

product_name	total_quantity	total_revenue
Smart ankle pants (wool like)	289	12931.48
Smooth jersey lined parka	199	9940.05
Ultra stretch active jogger pants	394	9830.30
Seamless down parka (3D cut)	48	7197.60
W's light souffle yarn L/S short cardigan	145	5792.75
W's ultra light down jacket	67	5691.65
W's pocketable UV protection parka	123	4913.85
Ultra stretch color jeans	121	4833.95
W's smart ankle pants	112	4474.40
Dry pique S/S polo shirt	176	4391.20



Unique Product Offerings

Q10) How many unique products were sold within the period of January 1, 2021, to December 31, 2022?

Analysis:

The output displays 153 unique products that were sold over a year. Businesses can strategize their marketing ad campaigns, optimize their product portfolio, and differentiate businesses from competitors.

```
84    #10. How many unique products were sold within the time period of January 1, 2021 to December 31, 2022?
85 •  SELECT COUNT(DISTINCT t.product_id) AS num_products_sold
86    FROM Transaction t INNER JOIN Product p USING (product_id)
87    WHERE t.transaction_date BETWEEN '2021-01-01' AND '2022-12-31';
```

100% 24:80

Result Grid | Filter Rows: Search | Export:

num_products_s...
153

Financial Performance

Q11) What was the total revenue generated from these product sales?

Analysis:

The output displays the decent health of the business. They can focus on making informed decisions to expand or strategize their product portfolio or any other major business decisions.

```
87    #11. What was the total revenue generated from these product sales?
88 •  SELECT ROUND(SUM(t.total_amount),2) AS total_revenue
89    FROM Transaction t
90    WHERE t.transaction_date BETWEEN '2021-01-01' AND '2022-12-31';
```

100% 64:90

Result Grid | Filter Rows: Search | Export:

total_revenue
202075.91

Effectiveness of Sales Channel

Q12) What is the distribution of transactions across different channels?

Analysis:

The output shows the maximum number of payment transactions through the website, application, and in-store. This analysis can help evaluate each channel's effectiveness, and decisions about investing in technologies for seamless payment processing across each platform can be made.

```
92    #12. What is the distribution of transactions across different channels?
93    SELECT c.channel, COUNT(t.transaction_id) AS num_transactions
94    FROM Channel c
95    LEFT JOIN Transaction t ON c.channel_id = t.channel_id
96    GROUP BY c.channel;
```

100% 64:90

Result Grid | Filter Rows: Search | Export:

channel	num_transactions
Application	335
In-store	316
Website	349

Top Categories with the Largest Offerings

Q13) What are the top 5 categories with the highest number of products?

Analysis:

The result shows the top 5 categories with the highest number of products. This will help businesses aim to improve their resource allocations on these products.

```
98    # 13. What are the top 5 categories with the highest number of products?
99 •  SELECT c.category_name, COUNT(p.product_id) as product_count
100   FROM Category c
101   LEFT JOIN Product p
102   ON c.category_id = p.category_id
103   GROUP BY c.category_id
104   ORDER BY product_count DESC
105   LIMIT 5;
```

100% 9:105

Result Grid | Filter Rows: Search | Export: | Fetch rows:

category_name	product_count
cut & sewn	369
inner & living	308
bottoms	202
outer	133
shirts	122

Business Conclusion

Conclusion

- **Revenue generated by each season:** Help identify the most profitable seasons and plan inventory and marketing strategies accordingly.
- **Average payment amount by payment method:** Help the company to understand customer preferences with respect to payment options. The company can plan to invest in technology in the most preferred payment options to further increase sales.
- **Revenue generated by each category:** Help identify the most popular product categories and making informed decisions about product assortment and pricing strategies.

Conclusion (cont.)

- **Revenue generated by each gender:** Help understand customer preferences by gender and tailor marketing efforts accordingly.
- **Customer transaction analysis:** This information can be used to target customer retention strategies and improve customer satisfaction. This can also help in targeted marketing initiatives.
- **Product sales analysis:** Help identify the best-performing products and making data-driven decisions about inventory management and product promotion strategies.

Lessons Learned

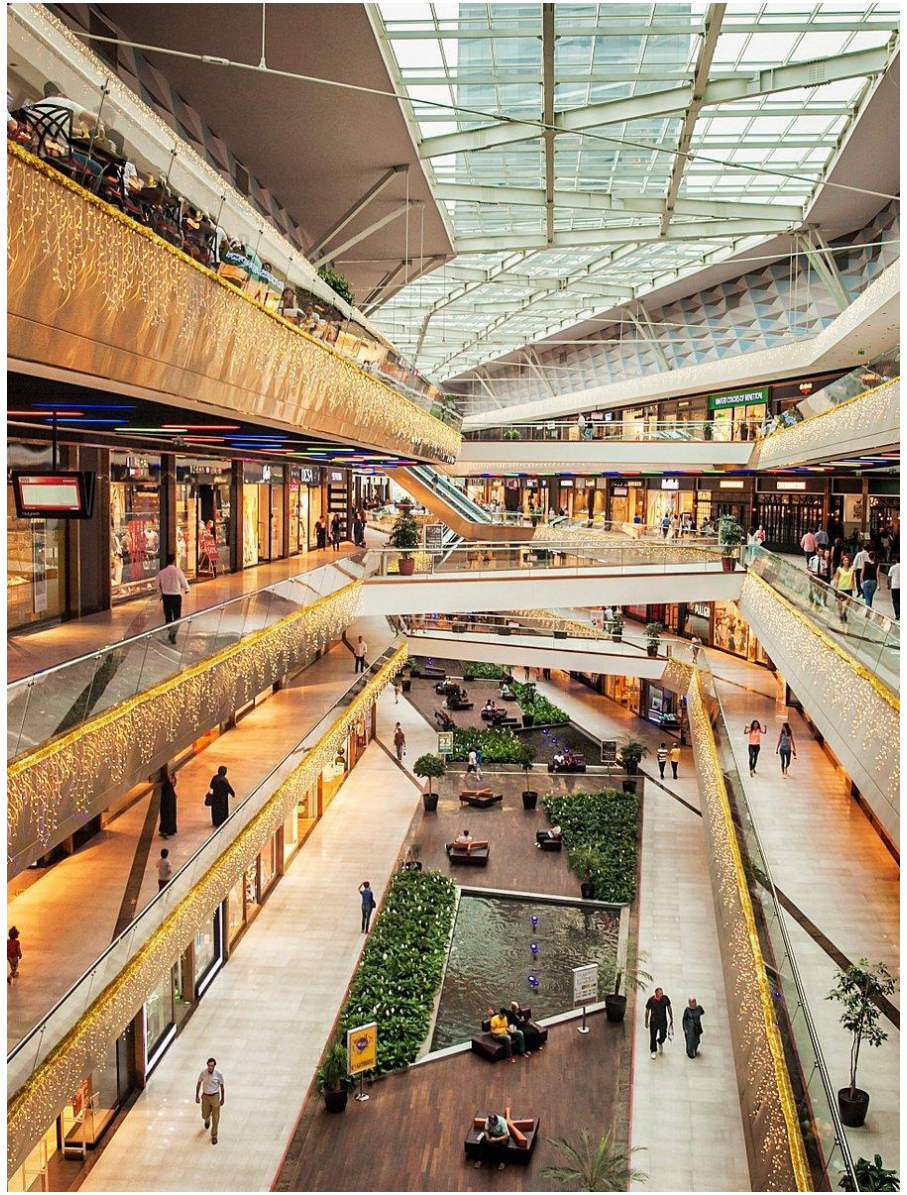
- **Proper database design**: Designing a well-structured and normalized database is crucial for efficient data management and analysis. Creating the tables and loading it to SQL.
- **Importance of data privacy and security:** Protecting sensitive customer and transaction data is critical to maintaining customer trust and complying with data protection regulations.
- **Data analysis and reporting:** Utilizing SQL queries and aggregate functions can help extract meaningful insights from the database for business decision-making.

Challenges

- **Data population:** It needed help at the beginning to populate the dataset to fulfill the business objectives.
- **Data quality:** Ensuring the accuracy and completeness of data in the database may require continuous data validation and cleaning processes.
- **Data volume**: In the current database, the data volume is low to run any trend analysis or any predictions.

Part 2: Data Analysis

- R
- Focusing on Retailing in Istanbul



Data Describing & Purpose of Analysis

DATA

- Analysis of shopping data in Istanbul.
- Data is gathered from 10 different shopping malls between 2021 and 2023.
- The dataset comprises of invoice numbers, customer IDs, age, gender, payment methods, product categories, quantity, price, order dates, and shopping mall locations.

PURPOSE

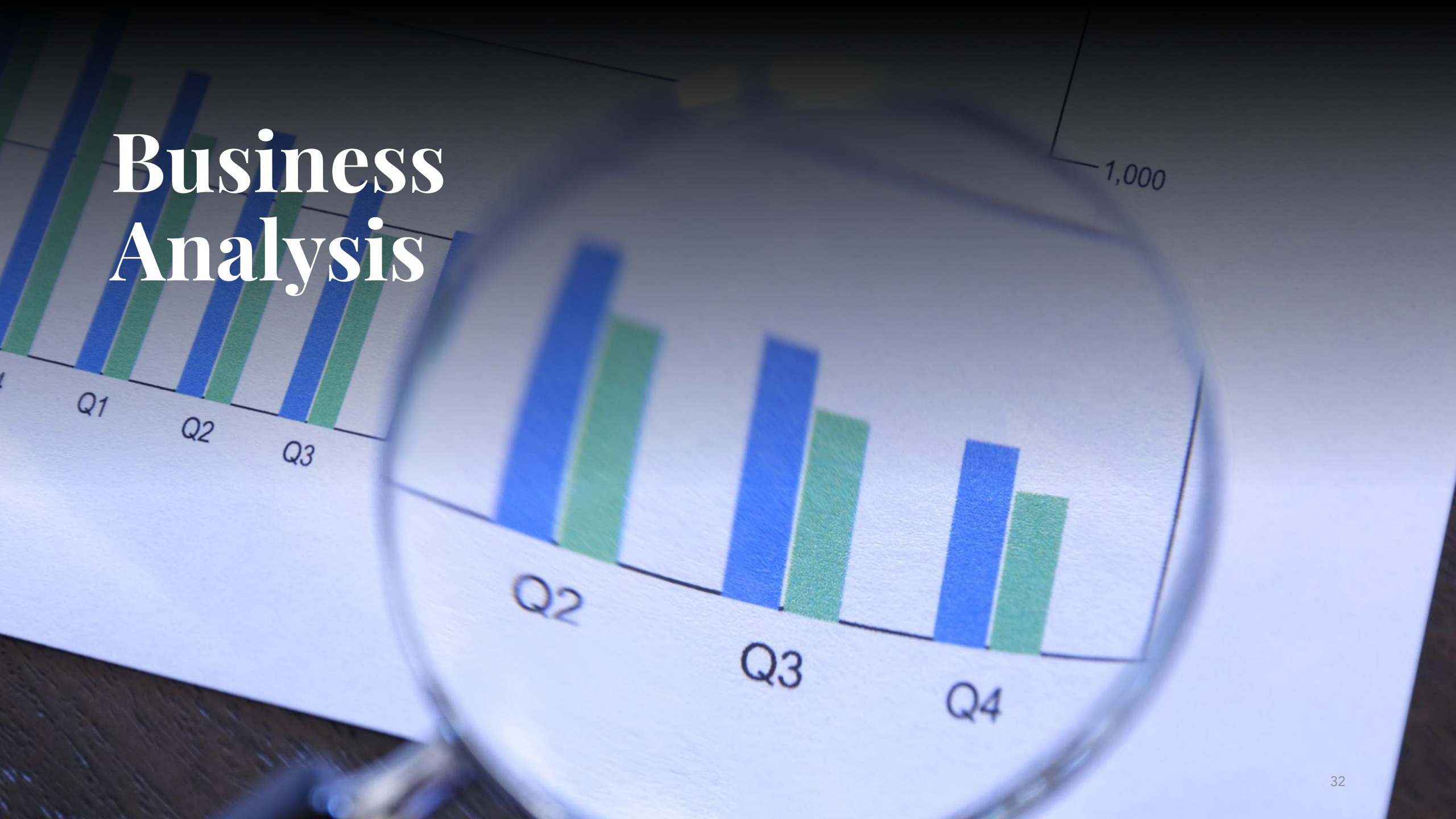
- The intended purpose of data analysis is to understand and identify the trends in customer purchase behavior, analyze which products and payment methods are most frequently used by the customers, which malls customers have visited the most, and how different customer demographics come into play while they shop.

Summary of Dataset

```
summary(customerDF)
```

```
##  shopping_mall       invoice_date         invoice_no        customer_id
##  Length:99457       Min.   :2021-01-01   Length:99457       Length:99457
##  Class :character   1st Qu.:2021-07-19   Class :character   Class :character
##  Mode  :character   Median :2022-02-05   Mode  :character   Mode  :character
##                     Mean   :2022-02-04
##                     3rd Qu.:2022-08-22
##                     Max.   :2023-03-08
##     gender                age         age_group           category
##  Length:99457       Min.   :18.00   Length:99457       Length:99457
##  Class :character   1st Qu.:30.00   Class :character   Class :character
##  Mode  :character   Median :43.00   Mode  :character   Mode  :character
##                     Mean   :43.43
##                     3rd Qu.:56.00
##                     Max.   :69.00
##     quantity         price            total_sale       payment_method
##  Min.   :1.000   Min.   :   5.23   Min.   :    5.23   Length:99457
##  1st Qu.:2.000   1st Qu.:  45.45   1st Qu.:  136.35   Class :character
##  Median :3.000   Median : 203.30   Median :  600.17   Mode  :character
##  Mean   :3.003   Mean   : 689.26   Mean   : 2528.79
##  3rd Qu.:4.000   3rd Qu.:1200.32   3rd Qu.: 2700.72
##  Max.   :5.000   Max.   :5250.00   Max.   :26250.00
```

Business Analysis



A. Descriptive Analysis

Summary statistics for numerical variables

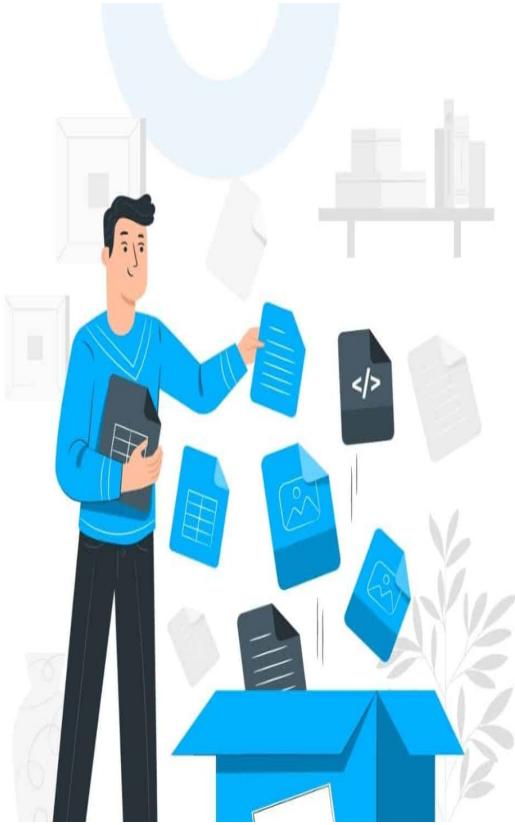
```
summary(customerDF[c("age", "quantity", "price")])
```

```
##       age           quantity         price
##  Min.   :18.00   Min.   :1.000   Min.   :   5.23
##  1st Qu.:30.00   1st Qu.:2.000   1st Qu.:  45.45
##  Median :43.00   Median :3.000   Median : 203.30
##  Mean   :43.43   Mean   :3.003   Mean   : 689.26
##  3rd Qu.:56.00   3rd Qu.:4.000   3rd Qu.:1200.32
##  Max.   :69.00   Max.   :5.000   Max.   :5250.00
```

Number of unique levels for categorical variables

```
sapply(customerDF[, c("shopping_mall", "gender", "category",
"payment_method", "age_group")], function(x) length(unique(x)))
```

```
##  shopping_mall         gender       category payment_method      age_group
##             10              2              8              3              6
```



CATEGORICAL DATA VS NUMERICAL DATA

Frequency table for categorical variables

```
table(customerDF$shopping_mall)
```

```
##
##       Cevahir AVM Emaar Square Mall    Forum Istanbul      Istinye Park
##              4991              4811              4947              9781
##            Kanyon  Mall of Istanbul         Metrocity      Metropol AVM
##             19823             19943             15011             10161
##    Viaport Outlet      Zorlu Center
##              4914              5075
```

```
table(customerDF$gender)
```

```
##
## Female   Male
##  59482  39975
```

```
table(customerDF$category)
```

```
##
##           Books        Clothing       Cosmetics Food & Beverage           Shoes
##            4981           34487           15097           14776           10034
##        Souvenir      Technology            Toys
##            4999            4996           10087
```

```
table(customerDF$payment_method)
```

```
##
##        Cash Credit Card  Debit Card
##       44447       34931       20079
```

```
table(customerDF$age_group)
```

```
##
##  10's  20's  30's  40's  50's  60's
##  3780 19263 19287 19153 18931 19043
```

```
Mode(customerDF$shopping_mall)
```

```
## [1] "Mall of Istanbul"
## attr(,"freq")
## [1] 19943
```

Measures of Variability

```
price_range <- range(customerDF$price, na.rm = TRUE)
cat("Price range: ", price_range[1], " - ", price_range[2], "\n")

## Price range:  5.23  -  5250
```

Calculate standard deviation and variance of price

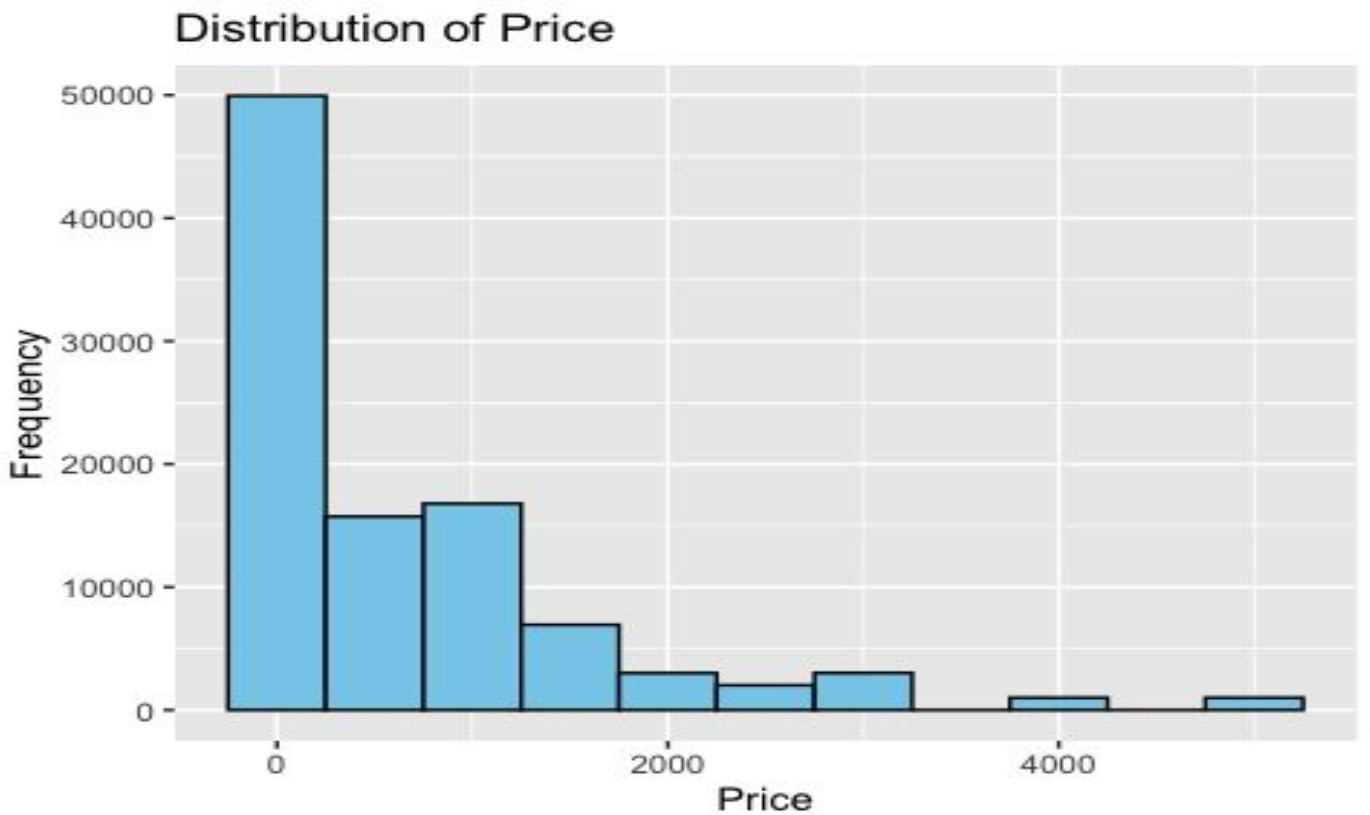
```
price_sd <- sd(customerDF$price, na.rm = TRUE)
cat("Price standard deviation: ", price_sd, "\n")

## Price standard deviation:  941.1846

price_var <- var(customerDF$price, na.rm = TRUE)
cat("Price variance: ", price_var, "\n")

## Price variance:  885828.4
```

Calculate interquartile range of price

```
price_iqr <- IQR(customerDF$price, na.rm = TRUE)
cat("Price interquartile range: ", price_iqr, "\n")

## Price interquartile range:  1154.87
```

Distribution

```
ggplot(customerDF, aes(x = price)) +
  geom_histogram(binwidth = 500, fill = "skyblue", color = "black") +
  ggtitle("Distribution of Price") +
  xlab("Price") +
  ylab("Frequency")
```

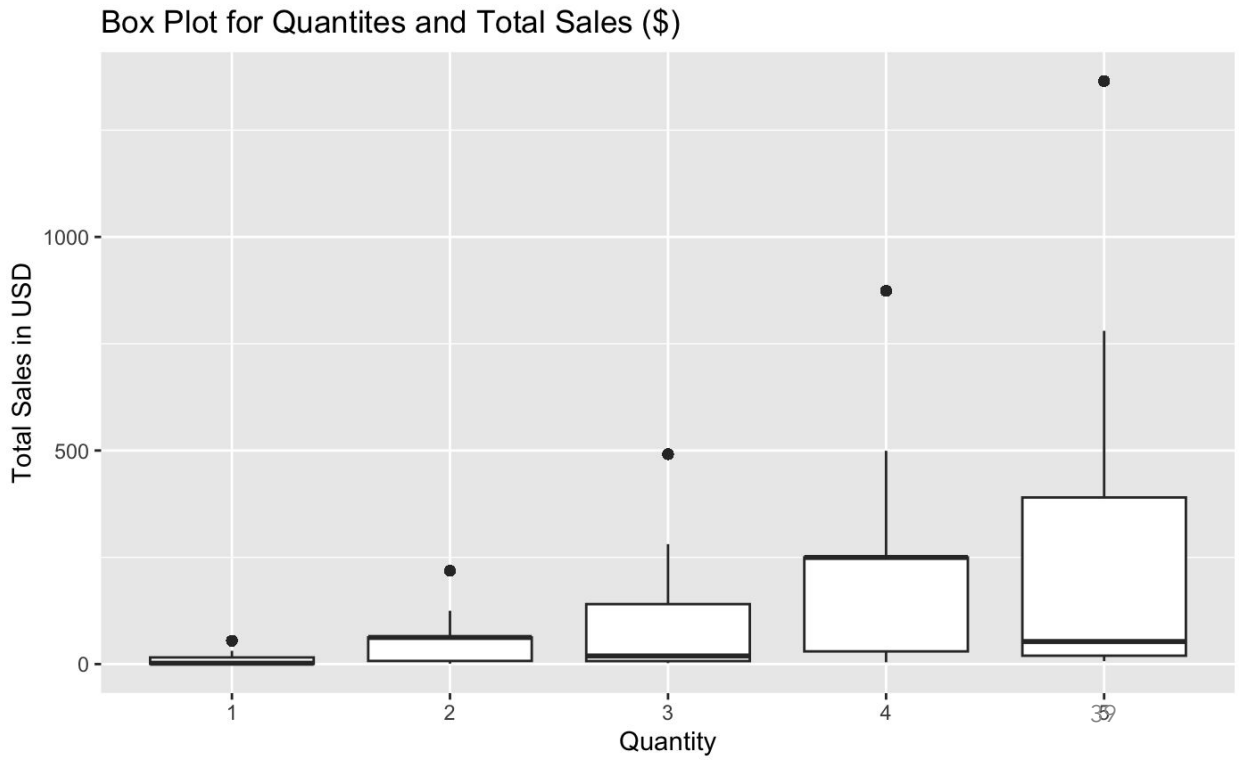


B. Boxplots

1. To understand the distribution pattern between quantities and total sales ($), box plots are set up with the x-axis as quantities and the y-axis as price in USD

As illustrated, it could be seen that the box in quantities - 5 (x = 5) has the most variety in the total sales with the highest outliner, while the box in quantities - 4 (x = 4) has the highest mean in total sales. It could be understood that customers prefer to spend more with a basket of 4 items, which could be related to expensive products - such as Technology. On the other hand, a basket of 5 items might vary throughout more affordable categories - such as Groceries.

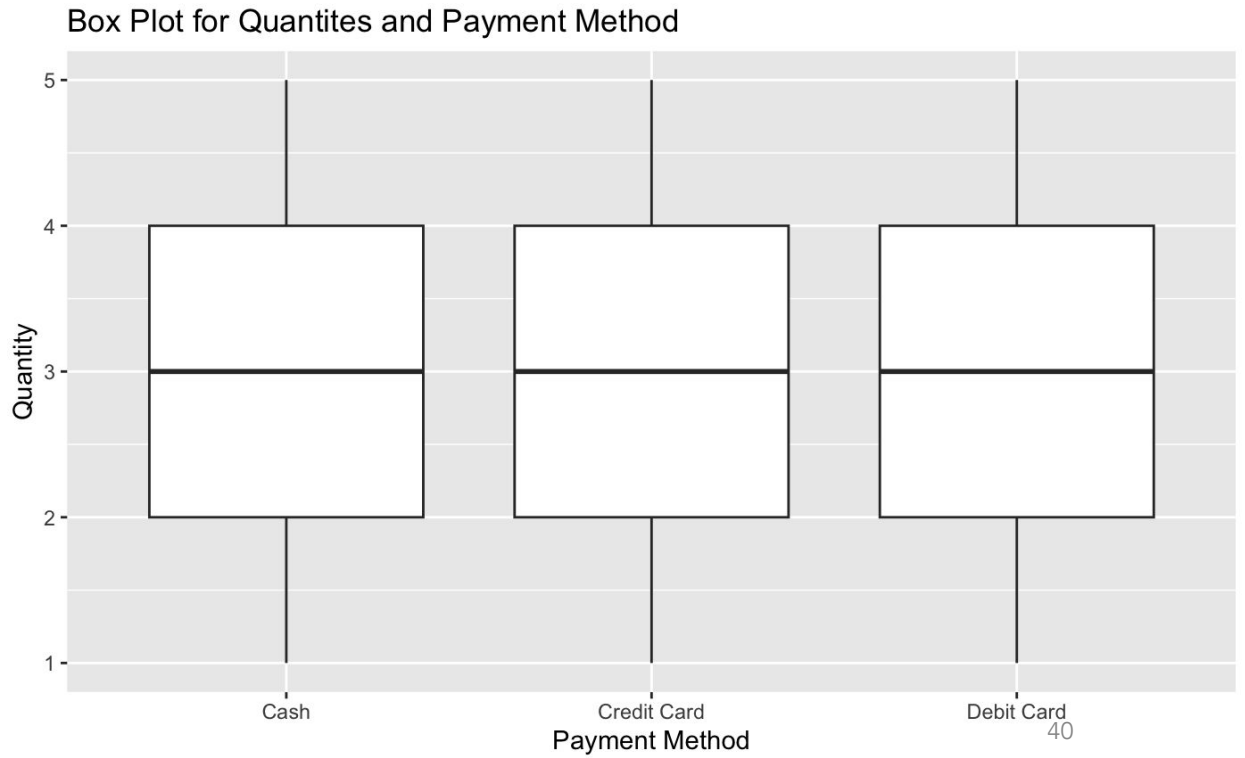
```
customerDF %>% ggplot(data = customerDF, mapping = aes(x =
factor(quantity), y = total_sale*0.052)) +
  geom_boxplot() +
  xlab("Quantity") +
  ylab("Total Sales in USD") +
  ggtitle("Box Plot for Quantites and Total Sales ($)")
```



2. To understand the distribution pattern between quantities and payment methods, box plots are set up with the x-axis as Payment Method (Cash/Credit Card/Debit Card) and the y-axis as quantities

As illustrated, it could be seen that the box in all 3 methods shared the same pattern with the average quantities that customers are willing to pay in each basket are 3 items, and the lowest amount for all 3 ways is 1 item per basket. The highest should be 5 units per basket, regardless of the payment method.

```
customerDF %>% ggplot(data = customerDF, mapping = aes(x =
factor(payment_method), y = quantity)) +
  geom_boxplot() +
  xlab("Payment Method") +
  ylab("Quantity") +
  ggtitle("Box Plot for Quantites and Payment Method")
```

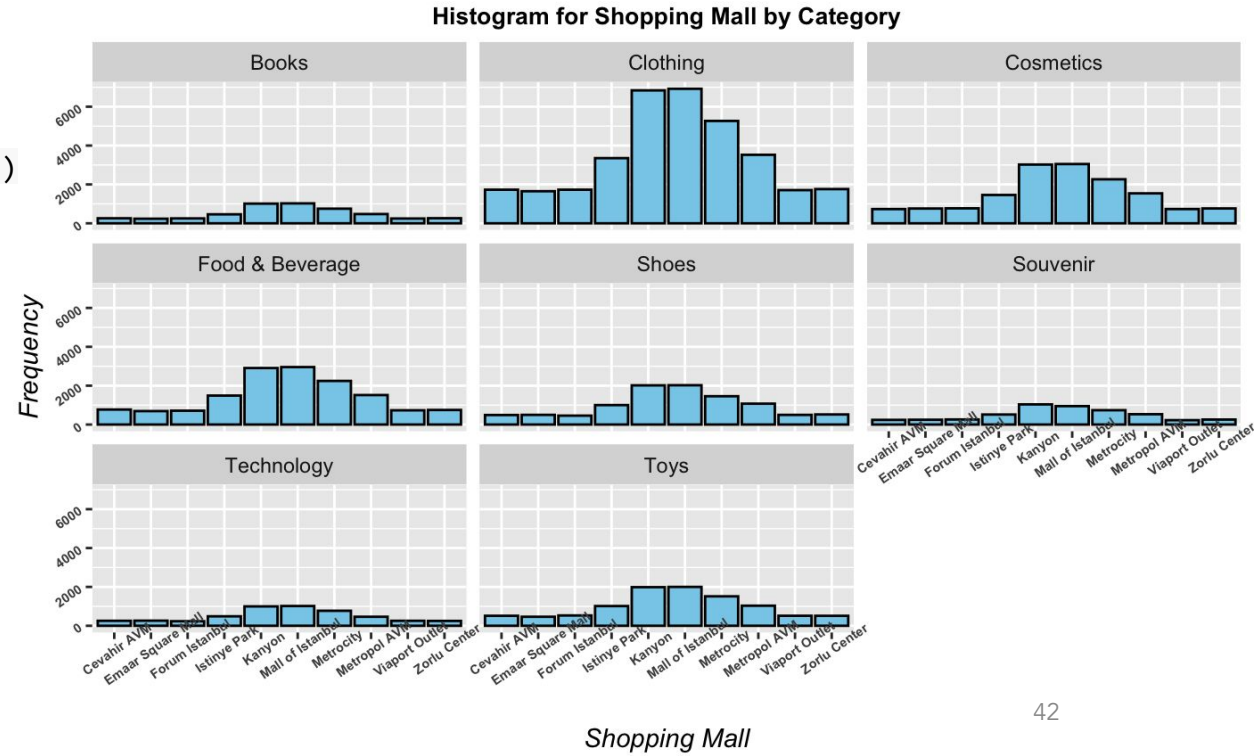


C. Histogram

1. To understand the purchasing pattern between categories per shopping mall, the histogram is set up with the x-axis as 10 Shopping Mall and the y-axis as Frequency in Purchasing Units, divided by Categories

As illustrated, the highest footfall shopping malls are Mall of Istanbul and Kanyon, in the center of Turkey. In addition, Clothing is the highest-selling category, followed by Cosmetics and Food & Beverage, which are all influenced by big fashion brands displayed in the Mall of Istanbul and Kanyon. Apart from that analysis, overall, the distribution pattern of shopping malls between categories is almost the same.

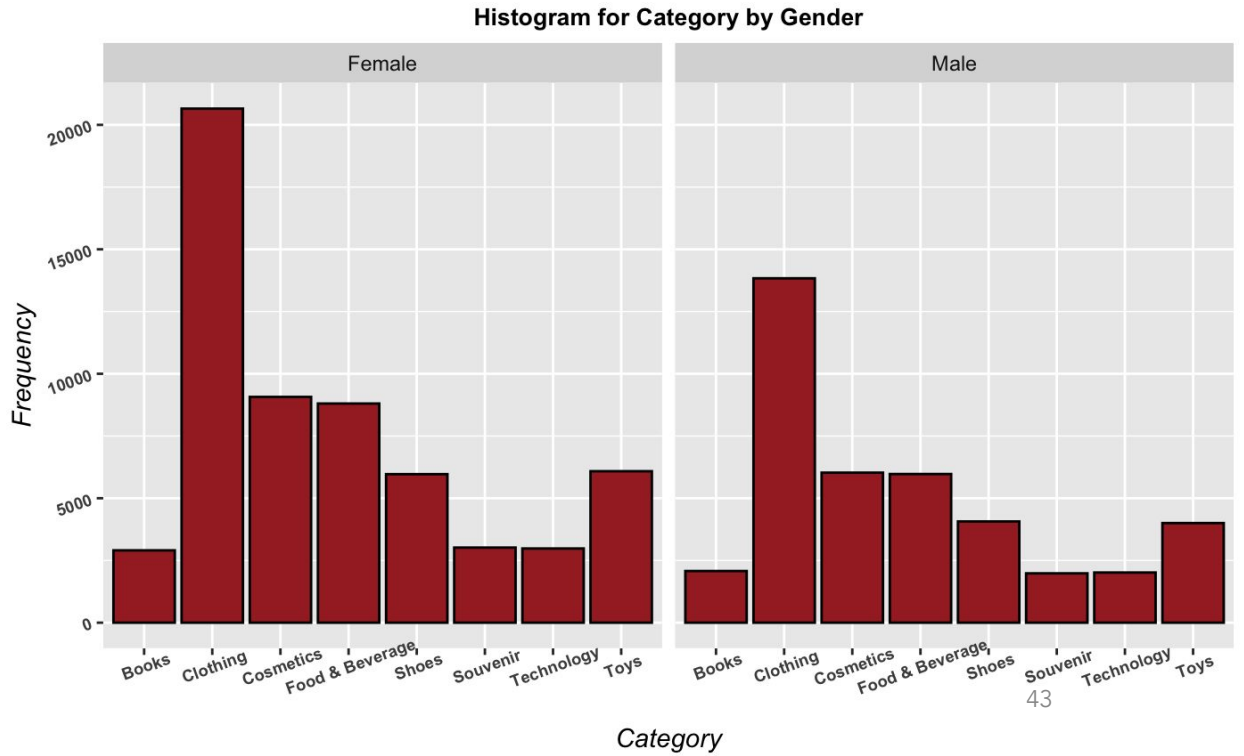
```
customerDF %>% ggplot(data = customerDF, mapping = aes(x = shopping_mall)) +
  geom_histogram(binwidth = 5, fill = 'skyblue', col = 'black', stat = "count")
+ facet_wrap(~ category) +
  xlab("Shopping Mall") +
  ylab("Frequency") +
  ggtitle("Histogram for Shopping Mall by Category") +
  theme(plot.title = element_text(face = 'bold', size = 10, hjust = .5),
        axis.title = element_text(face = 'italic'),
        axis.text = element_text(face = 'bold', size = 5.5, angle = 35))
```



2. To understand the purchasing pattern between categories by gender, the histogram is set up with the x-axis as Categories and the y-axis as Frequency in Purchasing Units, divided by Gender

As illustrated, the Female is the gender that goes shopping more frequently than the Male. In addition, similarly to the 1st histogram, Clothing is the highest-selling category, followed by Cosmetics and Food & Beverage, which can be predicted that Young Moms should be the main segment. Apart from that analysis, overall, the distribution pattern of genders between categories is almost the same.

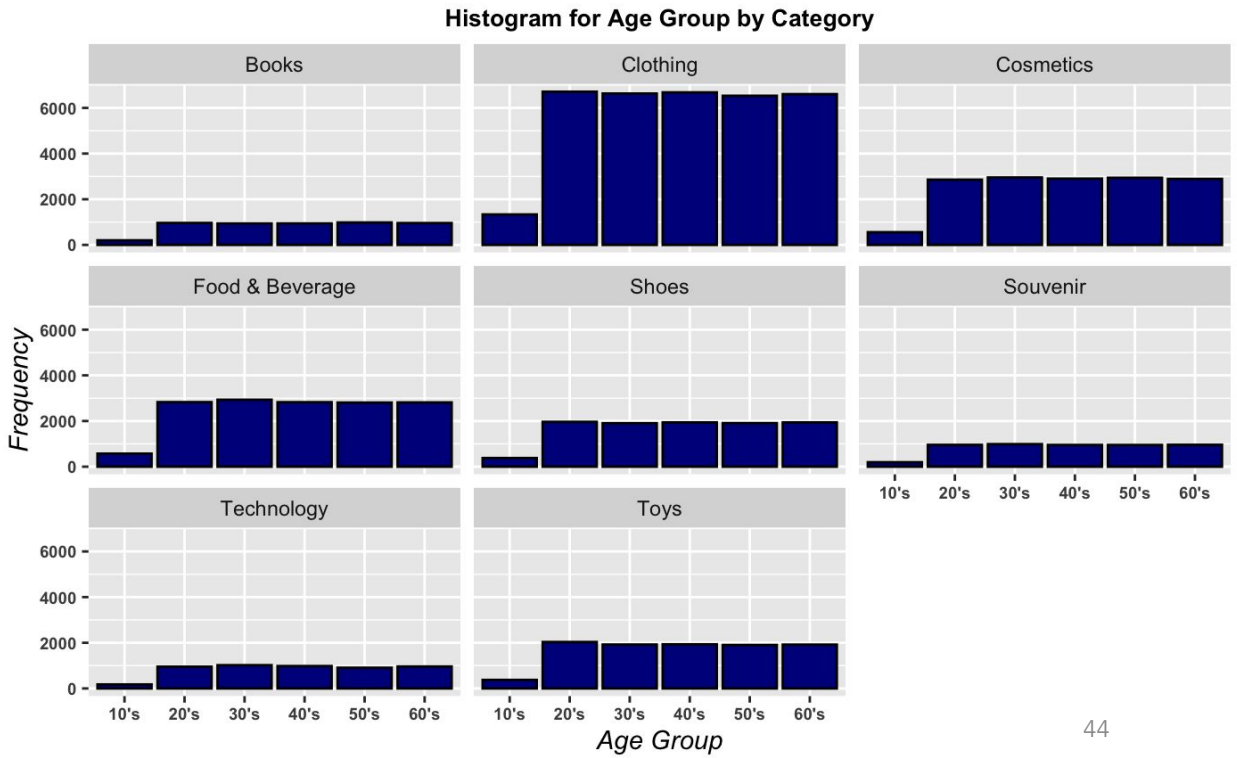
```
customerDF %>% ggplot(data = customerDF, mapping = aes(x = category)) +
  geom_histogram(binwidth = 5, fill = 'brown', col = 'black', stat =
"count") + facet_wrap(~ gender) +
  xlab("Category") +
  ylab("Frequency") +
  ggtitle("Histogram for Category by Gender") +
  theme(plot.title = element_text(face = 'bold', size = 10, hjust =
.5),
        axis.title = element_text(face = 'italic'),
        axis.text = element_text(face = 'bold', size = 7, angle = 20))
```



3. To understand the purchasing pattern between age groups by category, the histogram is set up with the x-axis as Age Group and the y-axis as Frequency in Purchasing Units, divided by Category

The distribution pattern of age groups between categories is almost the same. At the same time, age group 10's (below 20 years old) bought clothing and cosmetics more frequently than other categories; age group 20's (between 20 and 30 years old) tended to have a trend to purchase toys more than the different age groups, indicating that this group has more young children than the others.

```
customerDF %>% ggplot(data = customerDF, mapping = aes(x = age_group)) +
  geom_histogram(binwidth = 5, fill = 'darkblue', col = 'black', stat =
"count") + facet_wrap(~ category) +
  xlab("Age Group") +
  ylab("Frequency") +
  ggtitle("Histogram for Age Group by Category") +
  theme(plot.title = element_text(face = 'bold', size = 10, hjust = .5),
        axis.title = element_text(face = 'italic'),
        axis.text = element_text(face = 'bold', size = 7, angle = 0))
```

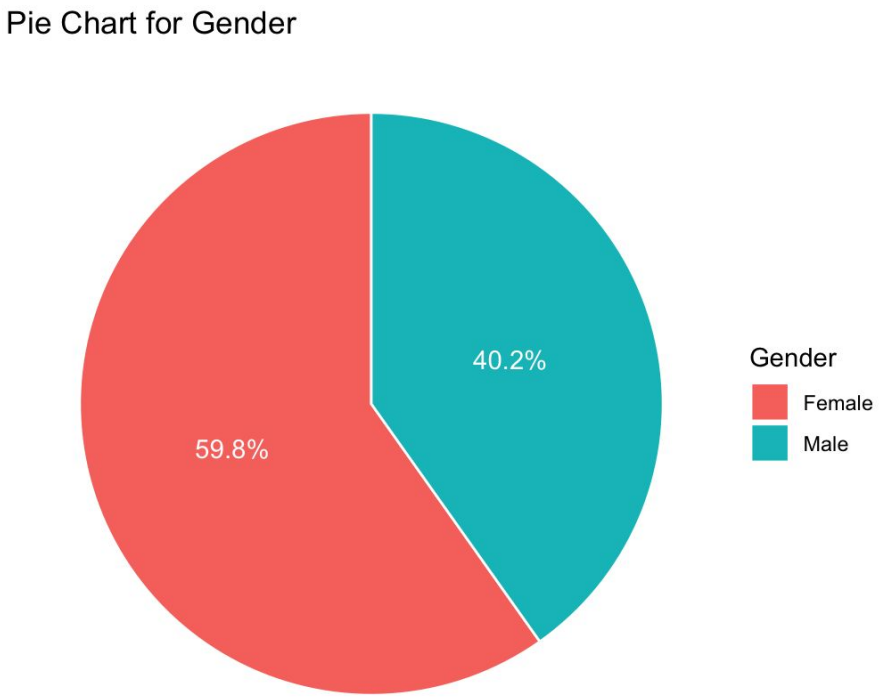


D. Pie Charts

1. Pie Chart for Gender

As expected, females in Istanbul have a stronger purchasing tendency than males. This can come from Young Single ones and Married Females who go shopping for the whole household.

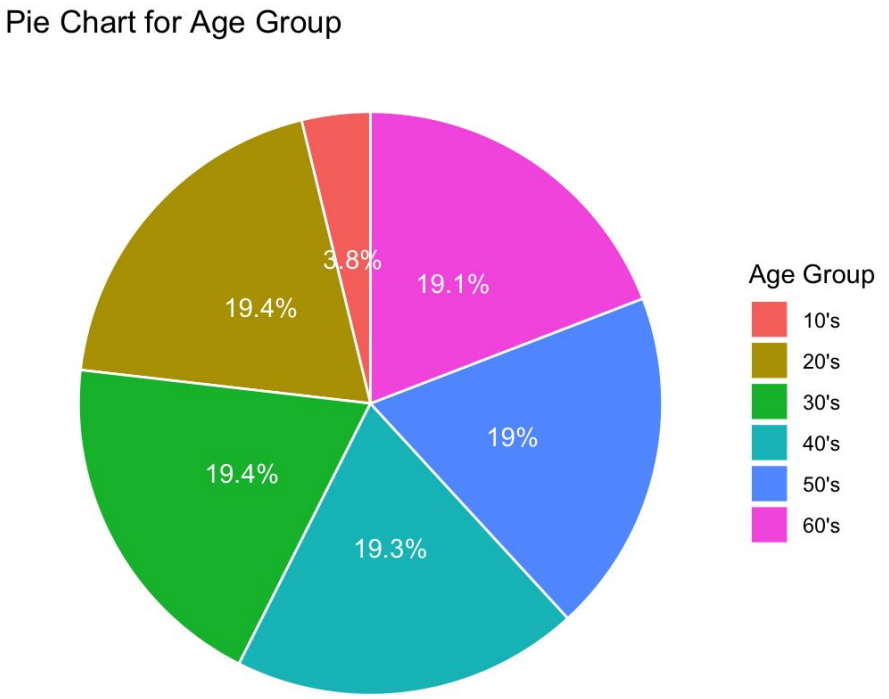
```
frequency_table4 <- customerDF %>%
  count(gender) %>%
  mutate(percentage = n / sum(n) * 100)

ggplot(frequency_table4, aes(x = "", y = percentage, fill = gender)) +
  geom_bar(stat = "identity", width = 1, color = "white") +
  coord_polar("y", start = 0) +
  geom_text(aes(label = paste0(round(percentage, 1), "%")), position =
position_stack(vjust = 0.5), color = "white") +
  ggtitle("Pie Chart for Gender") + labs(fill= "Gender") +
  theme_void()
```



2. Pie Chart for Age Group

Apart from the ‘10s segment (age below 20), each age group shared almost the same proportion with 19% per segment, yet the segment from the ‘20s to the ’30s, equally to customers age range from 20 to 40 years old, has a slightly higher percentage (19.4% each), indicating that they are the main customers to shopping malls.

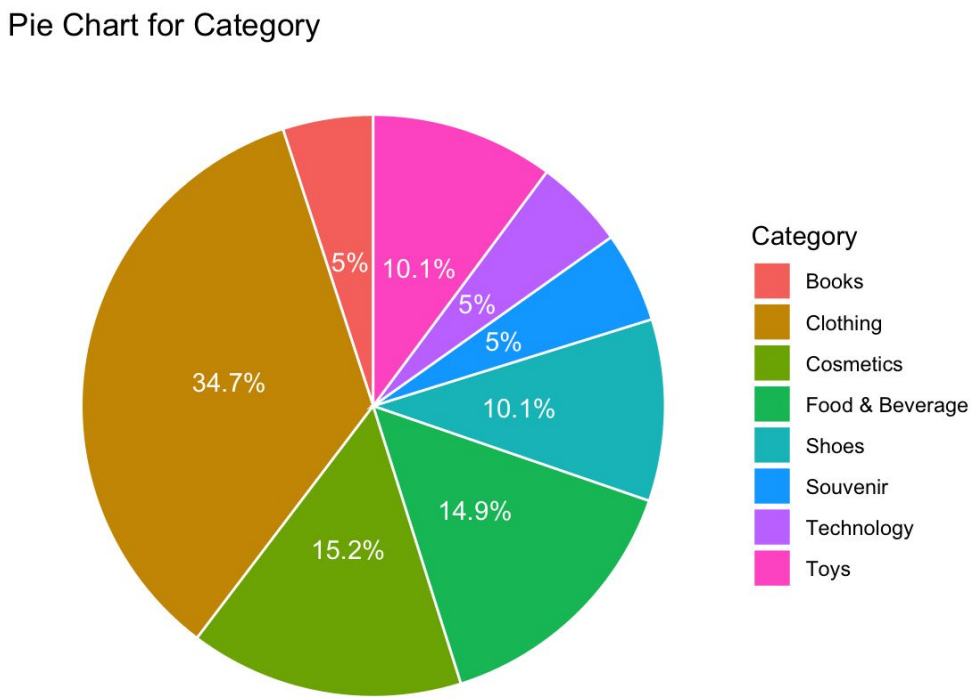
```
frequency_table2 <- customerDF %>%
  count(age_group) %>%
  mutate(percentage = n / sum(n) * 100)

ggplot(frequency_table2, aes(x = "", y = percentage, fill =
age_group)) +
  geom_bar(stat = "identity", width = 1, color = "white") +
  coord_polar("y", start = 0) +
  geom_text(aes(label = paste0(round(percentage, 1), "%")), position
= position_stack(vjust = 0.5), color = "white") +
  ggtitle("Pie Chart for Age Group") + labs(fill= "Age Group") +
  theme_void()
```



3. Pie Chart for Category

As analyzed above, Clothing is the highest category in Purchasing Units with 34.7%, followed by Cosmetics (15.2%) and Food & Beverage (14.9%). Shoes and Toys accounted for 10.1%, while the remaining categories shared 5% each.

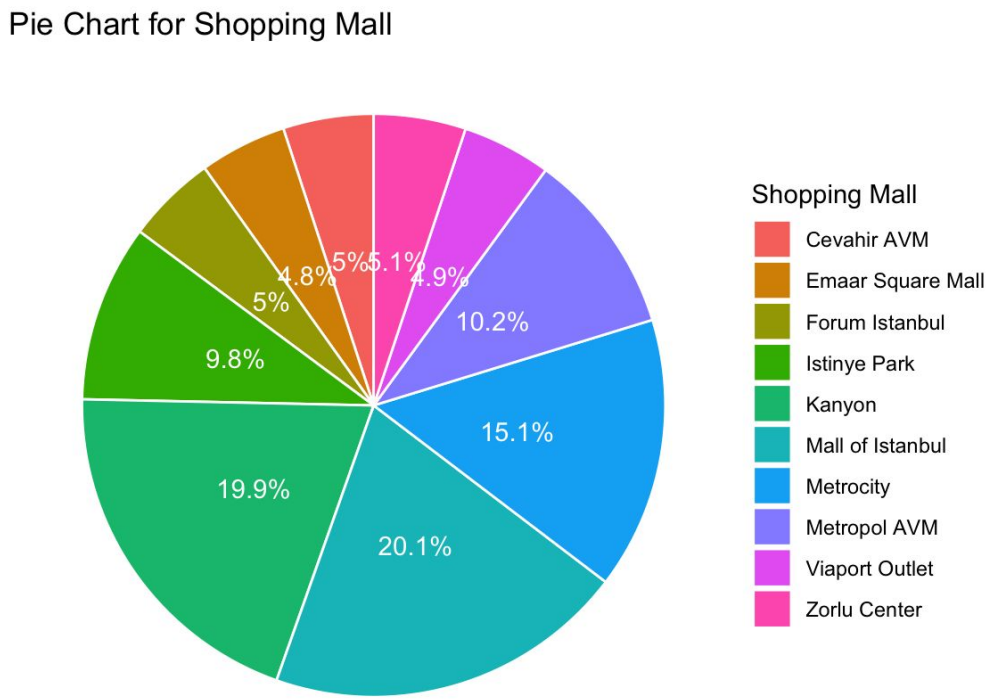
```
frequency_table <- customerDF %>%
  count(category) %>%
  mutate(percentage = n / sum(n) * 100)

ggplot(frequency_table, aes(x = "", y = percentage, fill =
category)) +
  geom_bar(stat = "identity", width = 1, color = "white") +
  coord_polar("y", start = 0) +
  geom_text(aes(label = paste0(round(percentage, 1), "%")),
position = position_stack(vjust = 0.5), color = "white") +
  ggtitle("Pie Chart for Category") + labs(fill= "Category") +
theme_void()
```



4. Pie Chart for Shopping Malls

As mentioned above, the Mall of Istanbul and Kanyon are the two highest purchasing traffic, with 20.1% and 19.9%, respectively. Among 10 shopping malls, Mall of Istanbul, Kanyon, and Metrocity are the leading players in the retail business, with more than a 50% accumulated composition ratio. In comparison, the rest 7 malls compete with more minor positions in the industry.

```
frequency_table3 <- customerDF %>%
  count(shopping_mall) %>%
  mutate(percentage = n / sum(n) * 100)

ggplot(frequency_table3, aes(x = "", y = percentage, fill = shopping_mall))
+
  geom_bar(stat = "identity", width = 1, color = "white") +
  coord_polar("y", start = 0) +
  geom_text(aes(label = paste0(round(percentage, 1), "%")), position =
position_stack(vjust = 0.5), color = "white") +
  ggtitle("Pie Chart for Shopping Mall") + labs(fill= "Shopping Mall") +
theme_void()
```

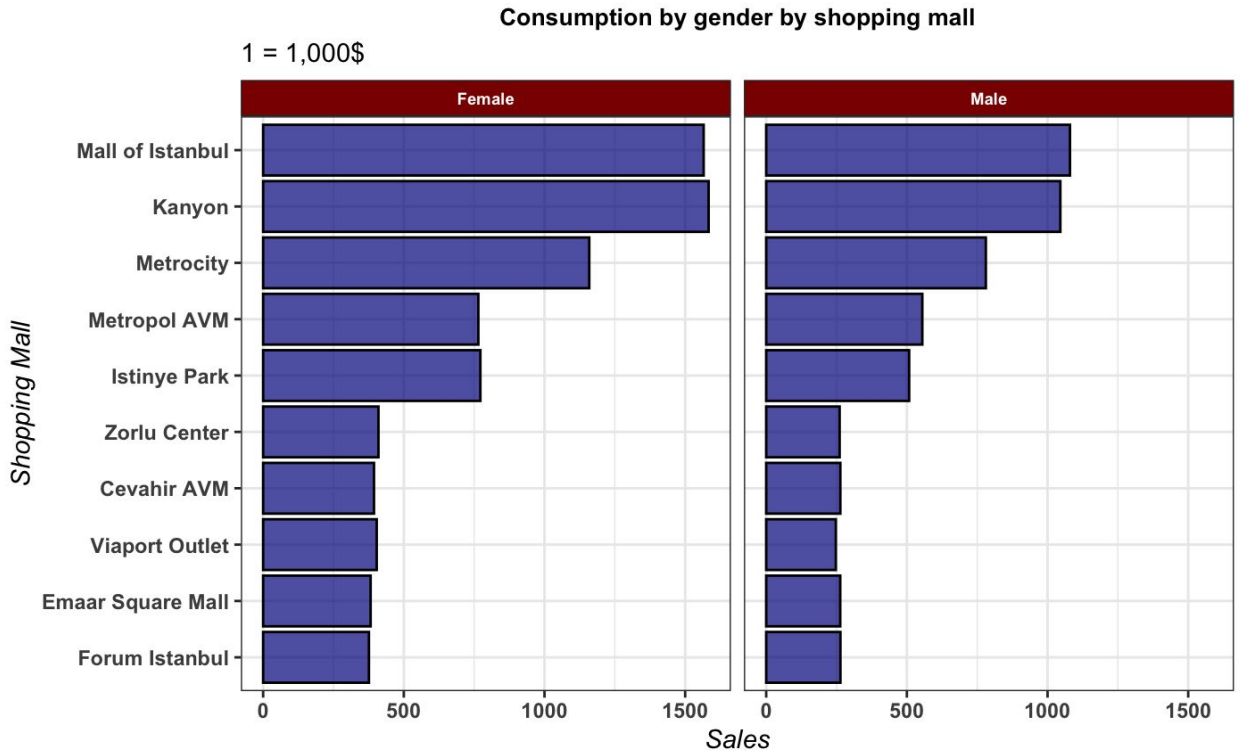


E. Sub-groups Faceting

1. By Gender and Shopping Mall

It can be easily seen that Female is the leading customer segment in Istanbul shopping, with 59,482 records - 1.48 times higher than Male.

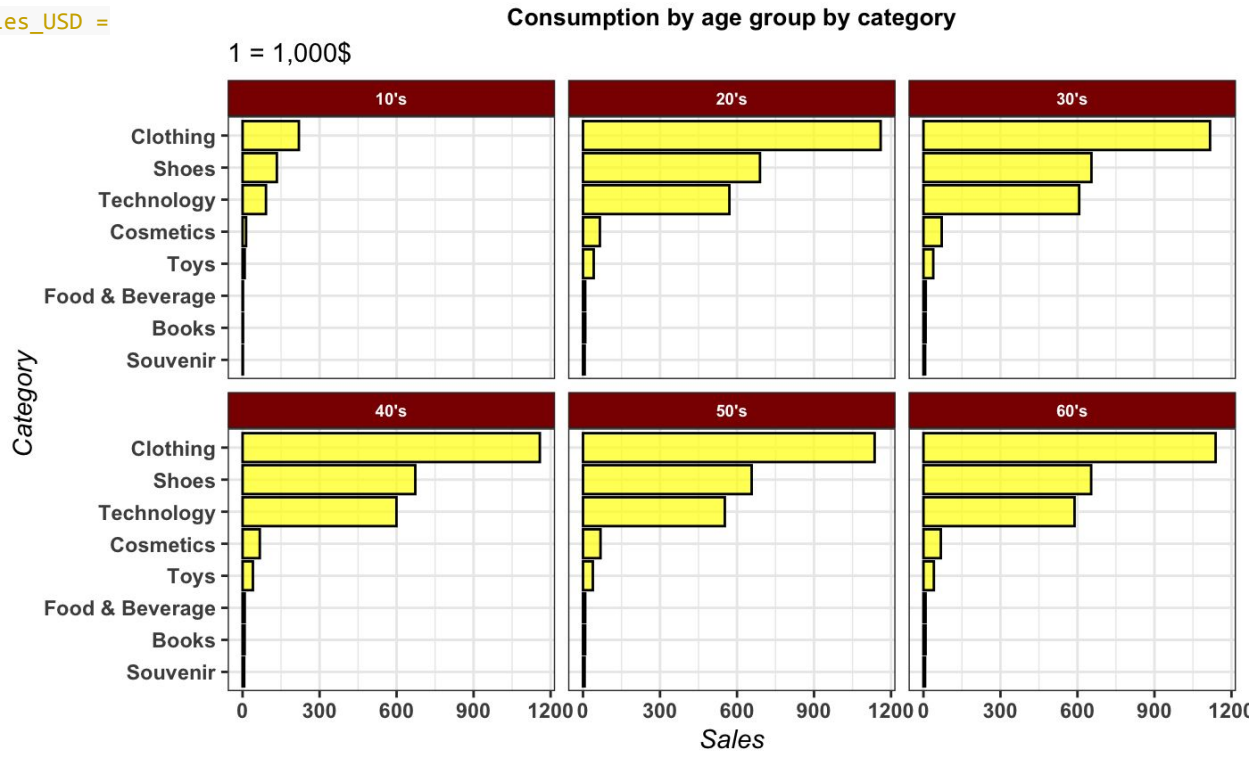
```
gender_vector <- customerDF %>% group_by(gender, shopping_mall) %>%
summarise(customer = n(), sales_USD = sum(total_sale*0.052/1000), .groups = "drop")
%>% arrange(sales_USD %>% desc())

gender_vector %>%
   ggplot(aes(reorder(shopping_mall, sales_USD), sales_USD)) +
  geom_col(position = 'dodge', fill = 'darkblue', col = 'black', alpha = .7) +
  coord_flip() +
  facet_wrap(~ gender) +
  labs(title = 'Consumption by gender by shopping mall',
       subtitle = '1 = 1,000$') +
  ylab('Sales') +
  xlab('Shopping Mall') +
  theme_bw() +
  theme(plot.title = element_text(face = 'bold', size = 10, hjust = .5),
        axis.title.y = element_text(face = 'italic'),
        axis.title.x = element_text(face = 'italic'),
        axis.text.x = element_text(face = 'bold'),
        axis.text.y = element_text(face = 'bold'),
        strip.text.x = element_text(face = 'bold' , color = 'white', size = 7),
        strip.background.x = element_rect(fill = 'darkred'))
```



2. By Age and Category

Except for age group 10's, the age range between the '20s and '60s is the target in shopping, whereas customers from 40 to 50 years old have the most vital purchasing power (sales amount = $695,578) among 6 age groups. Customers living in Istanbul have the highest tendency in clothing shopping with 34,487 records - $1,615,935. At the same time, Technology is 3rd highest sales amount category while there are 4,996 purchasers, indicating customers invested money in high-tech products.

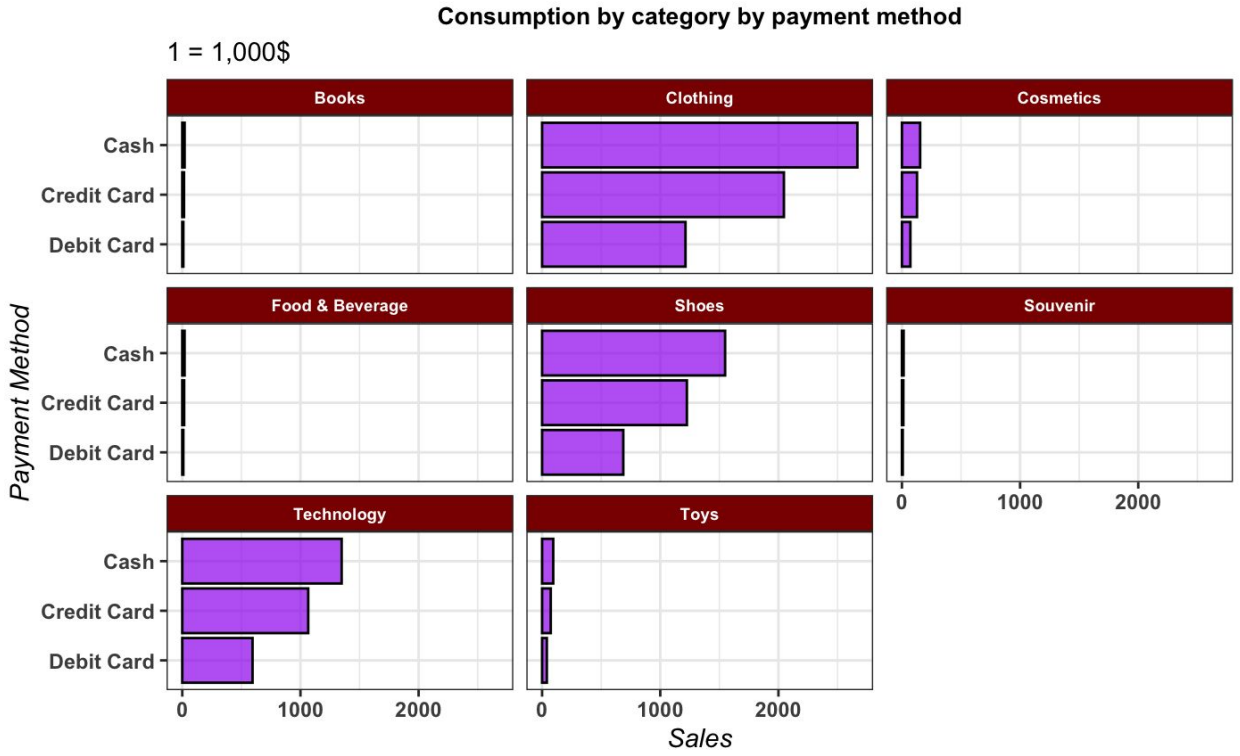
```
age_vector <- customerDF %>% group_by(age_group,category) %>% summarise(customer = n(), sales_USD =
sum(total_sale*0.052/1000), .groups = "drop") %>% arrange(sales_USD %>% desc())

age_vector %>%
  ggplot(aes(reorder(category, sales_USD), sales_USD)) +
  geom_col(position = 'dodge', fill = 'yellow', col = 'black', alpha = .7) +
  coord_flip() +
  facet_wrap(~ age_group) +
  labs(title = 'Consumption by age group by category',
       subtitle = '1 = 1,000$') +
  ylab('Sales') +
  xlab('Category') +
  theme_bw() +
  theme(plot.title = element_text(face = 'bold', size = 10, hjust = .5),
        axis.title.y = element_text(face = 'italic'),
        axis.title.x = element_text(face = 'italic'),
        axis.text.x = element_text(face = 'bold'),
        axis.text.y = element_text(face = 'bold'),
        strip.text.x = element_text(face = 'bold' , color = 'white', size = 7),
        strip.background.x = element_rect(fill = 'darkred'))
```



3. By Category by Payment Method

Customers living in Istanbul have the highest cash usage when purchasing, while the debit card is the weakest among the three methods. Besides that, the gap between cash and credit card use is the biggest in Clothing, whereas that could be smaller in Cosmetics, Technology, and Shoes.

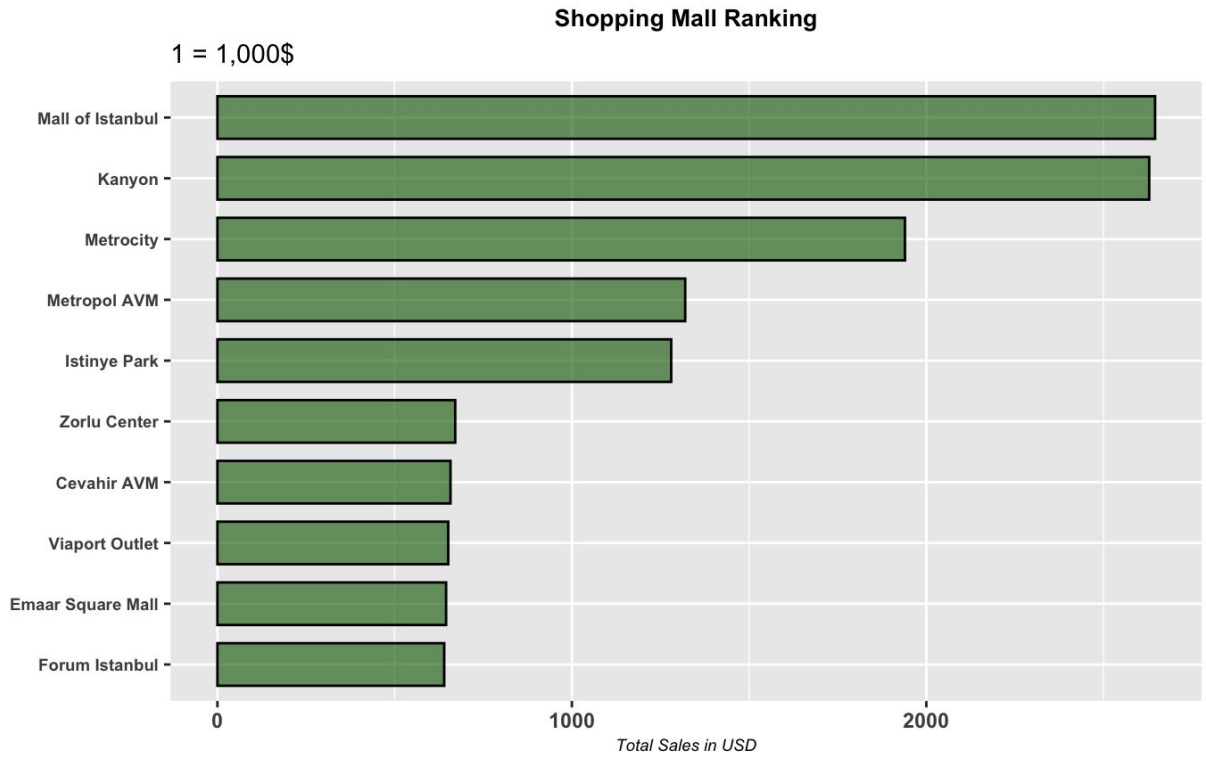
```
category_vector <- customerDF %>% group_by(category,payment_method) %>% summarise(customer =
n(), sales_USD = sum(total_sale*0.052/1000), .groups = "drop") %>% arrange(sales_USD %>%
desc())

category_vector %>%
  ggplot(aes(reorder(payment_method, sales_USD), sales_USD)) +
  geom_col(position = 'dodge', fill = 'purple', col = 'black', alpha = .7) +
  coord_flip() +
  facet_wrap(~ category) +
  labs(title = 'Consumption by category by payment method',
       subtitle = '1 = 1,000$') +
  ylab('Sales') +
  xlab('Payment Method') +
  theme_bw() +
  theme(plot.title = element_text(face = 'bold', size = 10, hjust = .5),
        axis.title.y = element_text(face = 'italic'),
        axis.title.x = element_text(face = 'italic'),
        axis.text.x = element_text(face = 'bold'),
        axis.text.y = element_text(face = 'bold'),
        strip.text.x = element_text(face = 'bold' , color = 'white', size = 7),
        strip.background.x = element_rect(fill = 'darkred'))
```



4. By Shopping Mall

The data set was collected from 10 malls in Istanbul. Among these, Mall of Istanbul and Kanyon are the two malls that have the highest traffic, reflecting 19,943 records and 19,823 records, respectively. There are 4 malls out of 10 that recorded more than 10,000 customers, representing the 4 most prominent malls in this city, while the other traffic range is between 4,800 and 10,000 customers.

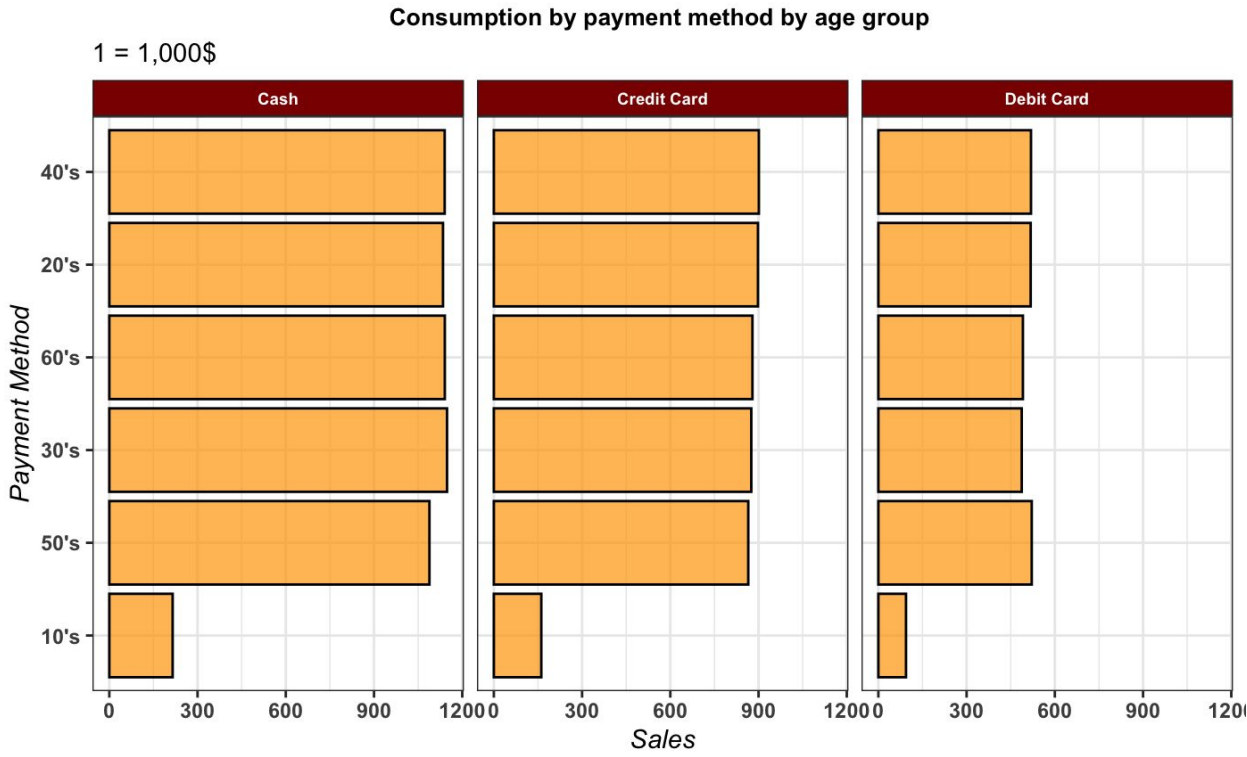
```
mall_vector <- customerDF %>% group_by(shopping_mall) %>% summarise(customer =
n(), sales_USD = sum(total_sale*0.052/1000), .groups = "drop") %>%
arrange(sales_USD %>% desc())

mall_vector %>%
  ggplot() +
  geom_col(aes(reorder(shopping_mall, sales_USD), sales_USD),
           fill = 'darkgreen', col = 'black', alpha = .6,width = .7) +
  coord_flip() +
  labs(title = 'Shopping Mall Ranking', subtitle = '1 = 1,000$') + ylab('Total
Sales in USD') +
  xlab('') +
  theme(plot.title = element_text(size = 10, face = 'bold', hjust = .5),
        axis.title.y = element_text(face = 'bold'),
        axis.text.y = element_text(face = 'bold', size = 7),
        axis.title.x = element_text(face = 'italic', size = 7),
        axis.text.x = element_text(face = 'bold'))
```



5. By Payment Method and Age Group

Cash is still the primary payment method, with the highest usage in both numbers of customers and sales amount.

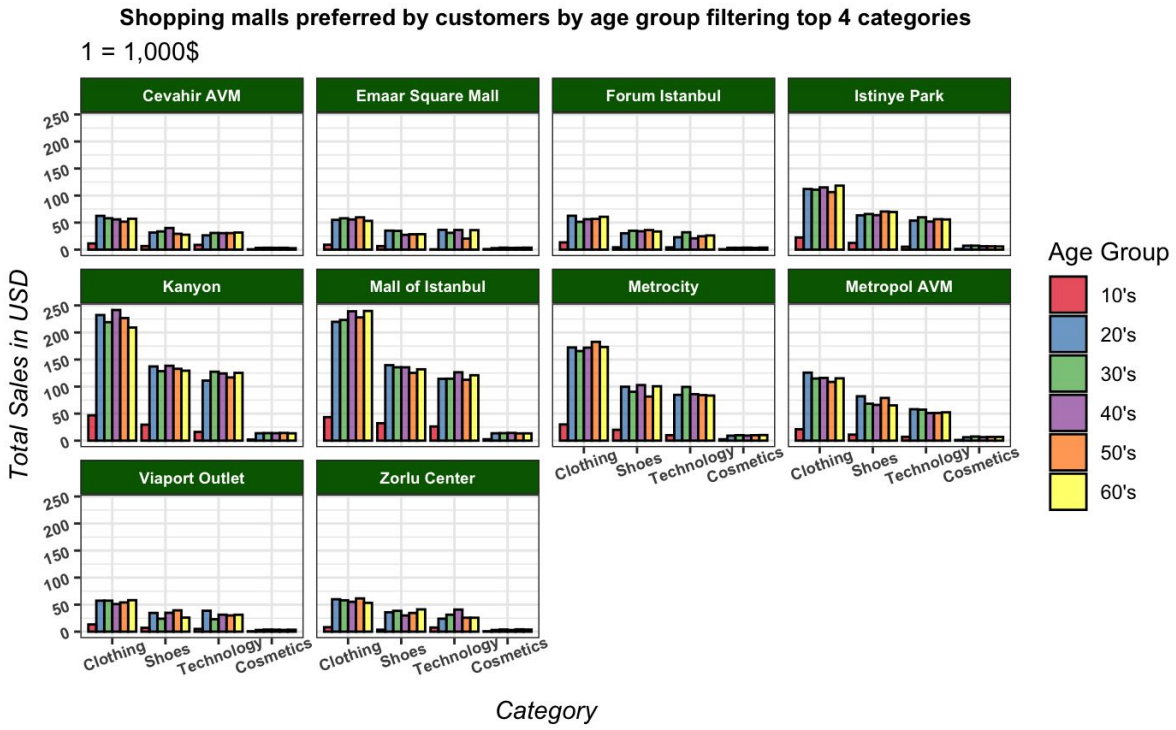
```
payment_vector <- customerDF %>% group_by(payment_method, age_group) %>%
summarise(customer = n(),sales_USD = sum(total_sale*0.052/1000), .groups = "drop")
%>% arrange(sales_USD %>% desc())

payment_vector %>%
  ggplot(aes(reorder(age_group, sales_USD), sales_USD)) +
  geom_col(position = 'dodge', fill = 'orange', col = 'black', alpha = .7) +
  coord_flip() +
  facet_wrap(~ payment_method) +
  labs(title = 'Consumption by payment method by age group ',
       subtitle = '1 = 1,000$') +
  ylab('Sales') +
  xlab('Payment Method') +
  theme_bw() +
  theme(plot.title = element_text(face = 'bold', size = 10, hjust = .5),
        axis.title.y = element_text(face = 'italic'),
        axis.title.x = element_text(face = 'italic'),
        axis.text.x = element_text(face = 'bold'),
        axis.text.y = element_text(face = 'bold'),
        strip.text.x = element_text(face = 'bold' , color = 'white', size = 7),
        strip.background.x = element_rect(fill = 'darkred'))
```



6. Preferred category by age group by shopping mall (high spending)

All 4 categories have a similar trend in every mall. In detail, the ‘40s (between 40 and 50 years old) customers and the ‘60s (more than 60 years old) purchase aggressively, followed by the ‘20s (between 20 and 30 years old) and the ‘50s (between 50 and 60 years old); yet the pattern varies between stores - for example: in Mall of Istanbul, the ‘40s and the ‘60s share the same power, while in Kanyon, the ‘40s has the more substantial purchasing power, which may be influenced by the demographic surrounding each mall.

```
customerDF %>% group_by(shopping_mall, age_group, category) %>%
  summarise(sales_USD = sum(total_sale*0.052/1000), .groups = "drop") %>%
  filter(category %in% c('Clothing','Shoes','Technology','Cosmetics')) %>%
  ggplot(aes(reorder(category, -sales_USD), sales_USD, fill = age_group)) +
  geom_col(position = 'dodge', col = 'black', alpha = .7) +
  facet_wrap(~ shopping_mall) +
  labs(title = 'Shopping malls preferred by customers by age group filtering top 4
categories',
       subtitle = '1 = 1,000$', fill = "Age Group")+
) +
  xlab('Category') + ylab('Total Sales in USD') + theme_bw() +
  theme(plot.title = element_text(face = 'bold', size = 10, hjust = .5),
        strip.background.x = element_rect(fill = 'darkgreen'),
        strip.text.x = element_text(face = 'bold', color = 'white', size = 7),
        axis.title = element_text(face = 'italic'),
        axis.text = element_text(face = 'bold', size = 7, angle = 20)) +
scale_fill_brewer(palette = 'Set1')
```

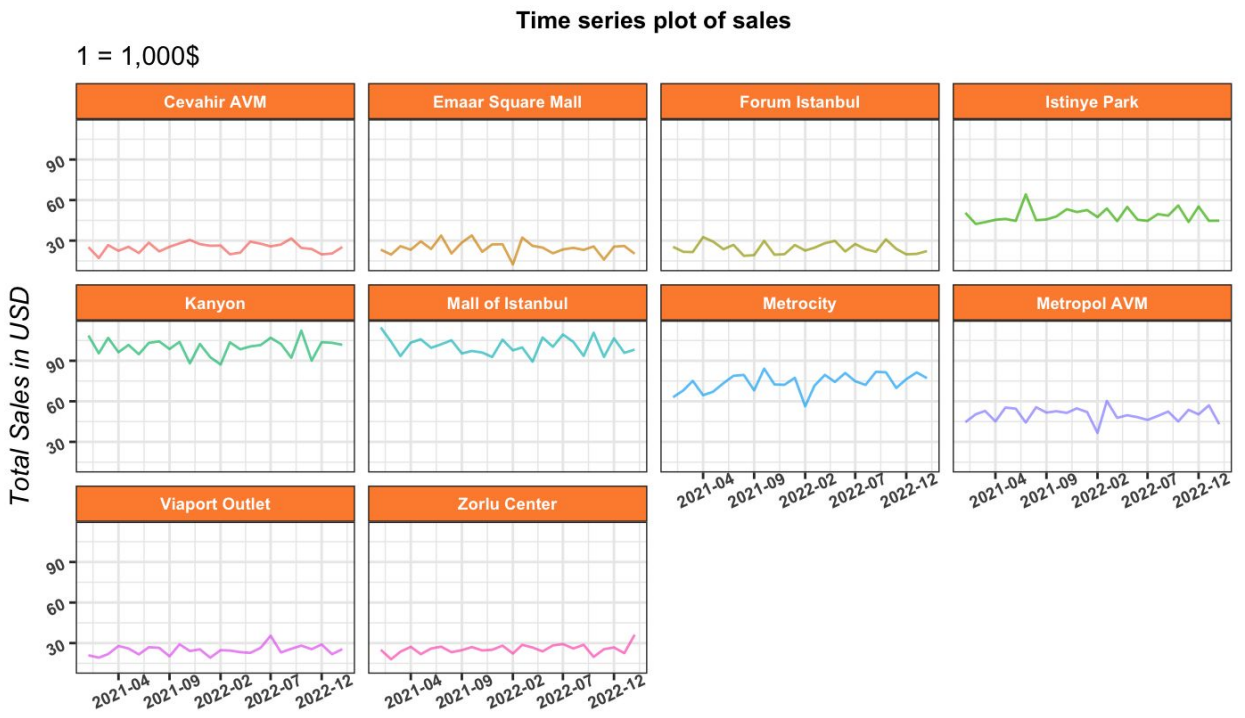


7. Preferred category by time series by shopping mall (high spending)

The other aspect to be analyzed is the time series to determine the busiest and slowest timing of the retailing industry. Based on this finding, managers can prepare an inventory and human resources to maximize profit and minimize the costs raised from the insufficient strategy.

Due to the most critical Turkish holidays allocated in these months, the busiest timing in the retailing industry should be between May and August, between November and January. The slowest timing, on the other hand, is between January and April, from September to October, since the holidays are not long enough to trigger customers' demands to go shopping.

```
customerDF %>% mutate(date2 = invoice_date %>% str_sub(1,7) %>% ym()) %>%
  group_by(date2, shopping_mall) %>%
  summarise(sales_USD = sum(total_sale*0.052/1000), .groups = "drop") %>%
  filter(date2 <= '2023-02-01') %>%
  ggplot(aes(date2, sales_USD, col = shopping_mall)) +
  geom_line(linewidth = .5, alpha = .7) +
  theme_bw() +
  facet_wrap(~ shopping_mall) +
  labs(title = 'Time series plot of sales',
       subtitle = '1 = 1,000$') +
  xlab('') + ylab('Total Sales in USD') +
  theme(plot.title = element_text(face = 'bold', size = 10, hjust = .5),
        strip.background = element_rect(fill = '#fd8d3c'),
        strip.text = element_text(face = 'bold',colour = 'white', size = 7),
        axis.title = element_text(face = 'italic'),
        axis.text = element_text(face = 'bold', size = 7, angle = 25),
        legend.position = 'none') +
  scale_x_date(date_breaks = '5 month', date_labels = '%Y-%m')
```

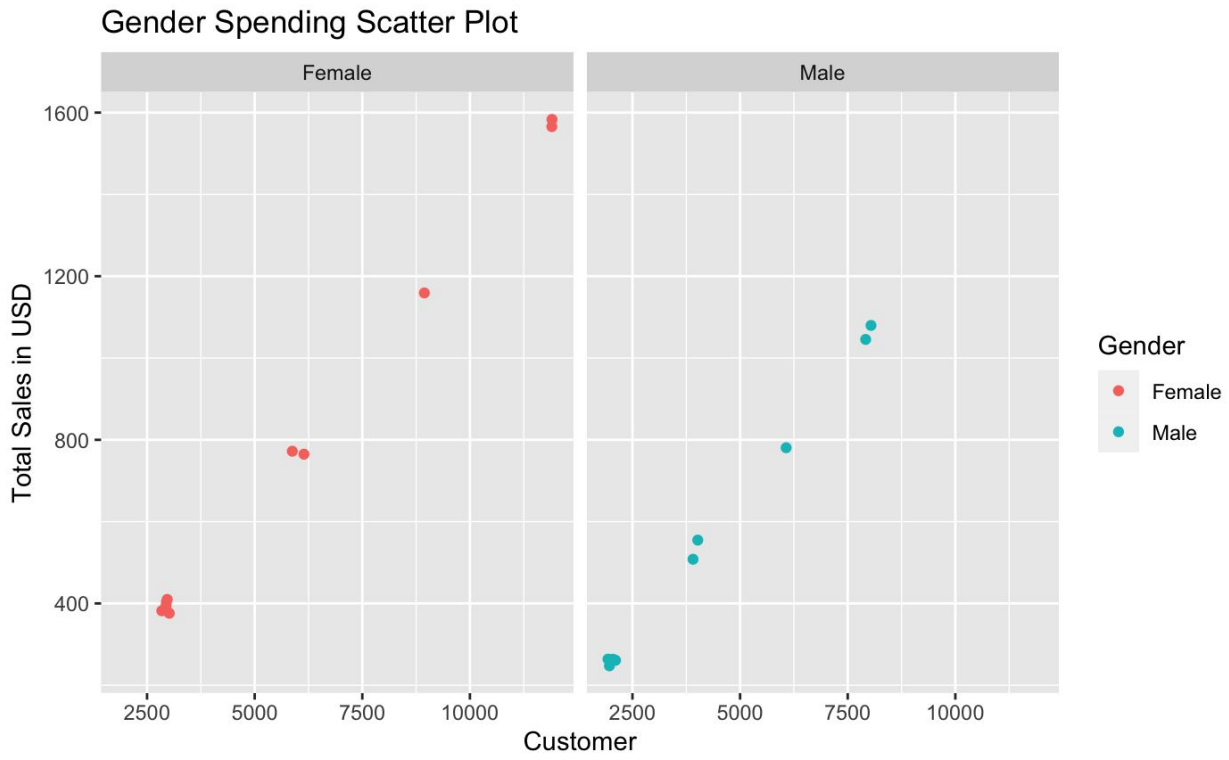


F. Scatter Plots

1. Gender Spending

As expected, females in Istanbul have a stronger purchasing tendency than males. This can come not only from Young Single ones but also from Married Females who go shopping for the whole household.

```
gender_vector %>%
  ggplot() + geom_point(aes(customer, sales_USD,
color=gender)) + facet_wrap(~gender) + labs(title =
"Gender Spending Scatter Plot") +  xlab('Customer') +
ylab('Total Sales in USD') + guides(color =
guide_legend(title = "Gender"))
```



2. Purchasing Behavior By Category

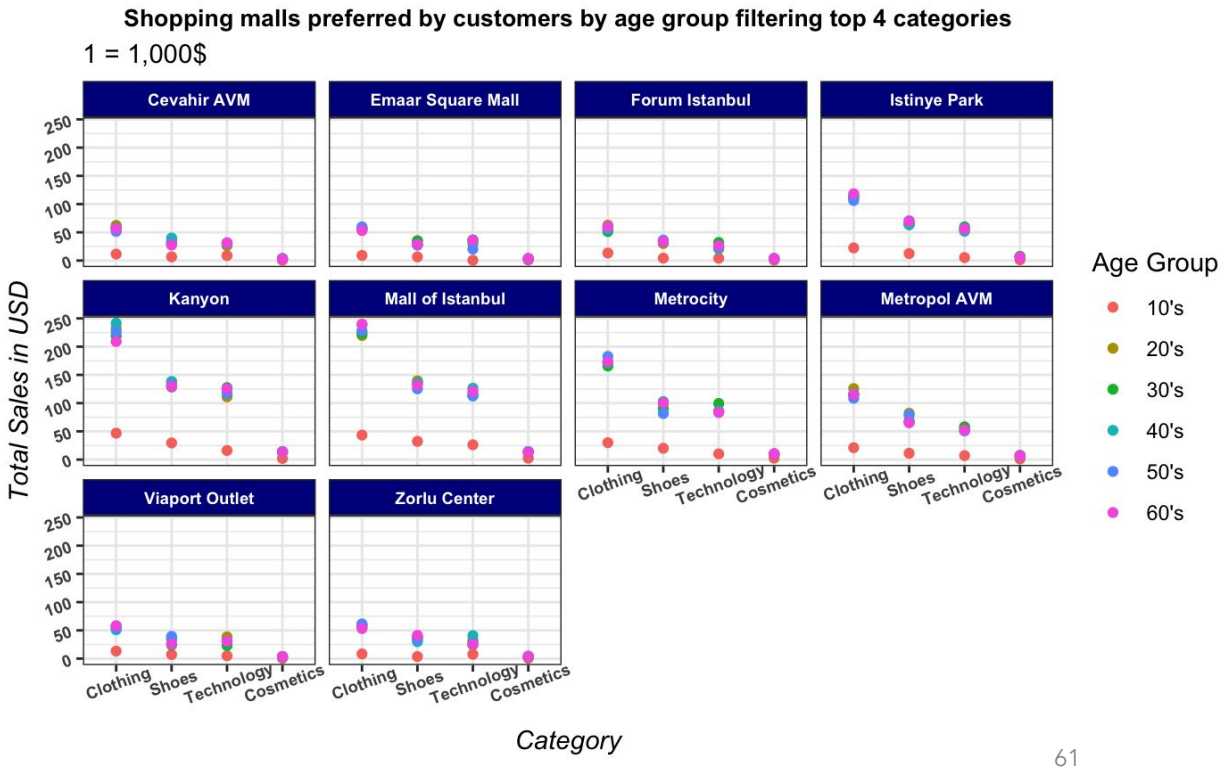
Cash is still the primary payment method, with the highest usage in both numbers of customers and sales amount.

```
customerDF %>% group_by(category, age_group, quantity,payment_method)
%>% summarise(sales_USD = sum(total_sale*0.052/1000), .groups = "drop")
%>%
  ggplot() + geom_point(aes(quantity, sales_USD, color=payment_method))
+ facet_wrap(. ~category) + labs(title = "Purchasing Behavior by
Category Scatter Plot",subtitle = '1 = 1,000$') + xlab('Quantity') +
ylab('Total Sales in USD') + guides(color = guide_legend(title =
"Payment Method")) + theme_bw() +
  theme(plot.title = element_text(face = 'bold', size = 10, hjust =
.5),
        strip.background.x = element_rect(fill = 'darkorange'),
        strip.text.x = element_text(face = 'bold', color = 'white',
size = 7),
        axis.title = element_text(face = 'italic'),
        axis.text.y = element_text(face = 'bold', size = 7, angle =
20)) +
scale_fill_brewer(palette = 'Set1')
```

3. Preferred category by age group with top 4 categories

All 4 categories have a similar trend in every mall. In detail, the ‘40s (between 40 and 50 years old) customers and the ‘60s (more than 60 years old) purchase aggressively, followed by the ‘20s (between 20 and 30 years old) and the ‘50s (between 50 and 60 years old); yet the pattern varies between stores - for example: in Mall of Istanbul, the ‘40s and the ‘60s share the same power, while in Kanyon, the ‘40s has the more substantial purchasing power, which may be influenced by the demographic surrounding each mall.

```
customerDF %>% group_by(shopping_mall, age_group, category) %>%
  summarise(sales_USD = sum(total_sale*0.052/1000), .groups = "drop") %>%
  filter(category %in% c('Clothing','Shoes','Technology','Cosmetics')) %>%
  ggplot(aes(reorder(category, -sales_USD), sales_USD, color = age_group)) +
  geom_point() +
  facet_wrap(~ shopping_mall) +
  labs(title = 'Shopping malls preferred by customers by age group filtering top 4
categories',
       subtitle = '1 = 1,000$') +
xlab('Category') + ylab('Total Sales in USD') + guides(color = guide_legend(title = "Age
Group")) +   theme_bw() +
  theme(plot.title = element_text(face = 'bold', size = 10, hjust = .5),
        strip.background.x = element_rect(fill = 'darkblue'),
        strip.text.x = element_text(face = 'bold', color = 'white', size = 7),
        axis.title = element_text(face = 'italic'),
        axis.text = element_text(face = 'bold', size = 7, angle = 20)) +
scale_fill_brewer(palette = 'Set1')
```



G. Regression

Hypothesis 1 - Age and Gender Impact Quantity of Purchase

Null Hypothesis (H0): Age and gender do not significantly impact purchase quantity.

Alternative Hypothesis (H1): Age and gender significantly impact purchase quantity.

Regression model 1:

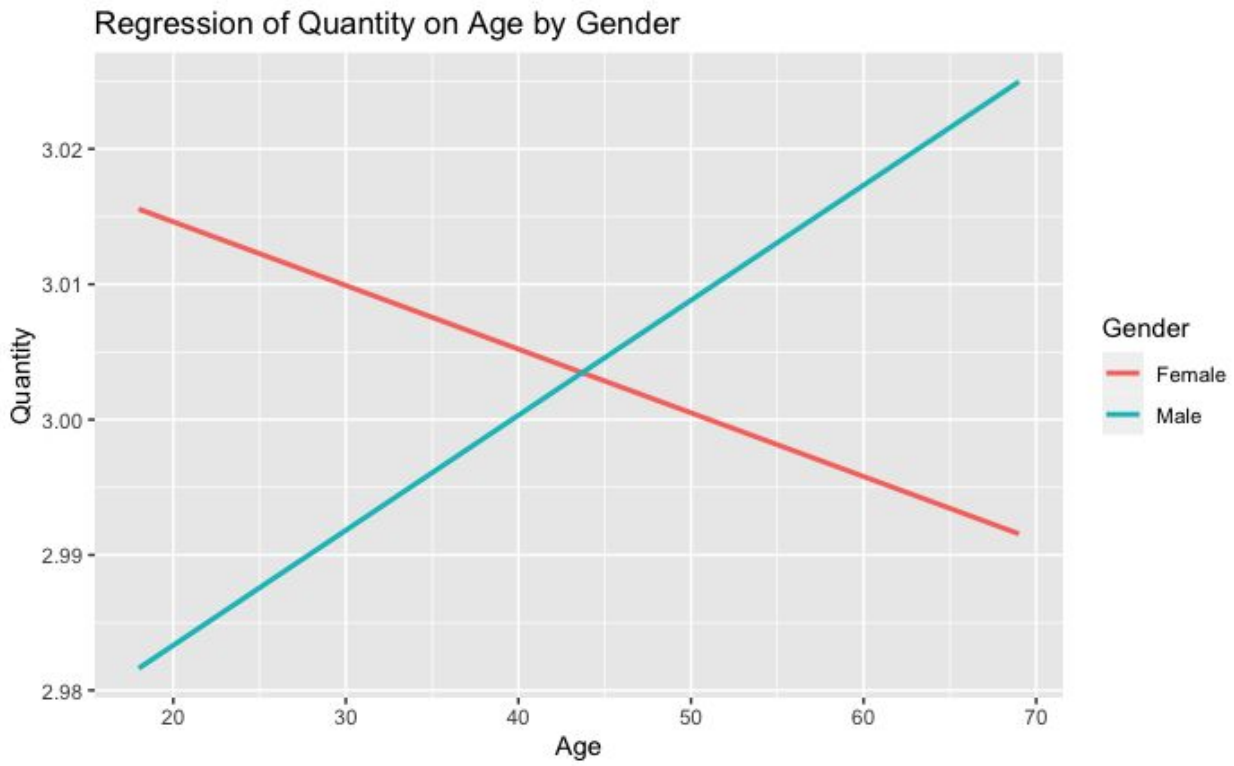
```
shopping_model1 <- lm(Quantity ~ Age + Gender, data = customerDF)
```

In this hypothesis, we are testing whether age and gender significantly impact the number of products purchased (Quantity) by customers. The regression model includes Age and Gender as predictor variables and Quantity as the response variable. The dataset contains the relevant variables and is used for analysis.

Output 1

Testing of Hypothesis 1: Age and Gender Impact Shopping Behavior

```
shopping_model1 <- lm(quantity ~ age + gender, data = customerDF)
summary(shopping_model1)
## Call:
## lm(formula = quantity ~ age + gender, data = customerDF)
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.00519 -1.00424 -0.00343  0.99739  1.99839
## Coefficients:
##Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.001e+00  1.422e-02 210.998   <2e-16 ***
## age          6.281e-05  2.989e-04   0.210    0.834
## genderMale  -3.748e-04  9.139e-03  -0.041    0.967
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.413 on 99454 degrees of freedom
## Multiple R-squared:  4.613e-07,  Adjusted R-squared:  -1.965e-05
## F-statistic: 0.02294 on 2 and 99454 DF,  p-value: 0.9773
```



Based on these results, we can conclude that Age and Gender do not significantly impact shopping behavior in this dataset, as they are not statistically significant predictors of Quantity. Based on the regression plot we see an inverse relationship between quantity and age for females, however the relationship is linear for males.

Hypothesis 2 - Payment Method Affects Purchase Amount

Null Hypothesis (H0): Payment method does not significantly affect the purchase amount.

Alternative Hypothesis (H1): The payment method significantly affects the purchase amount.

Regression model 2:

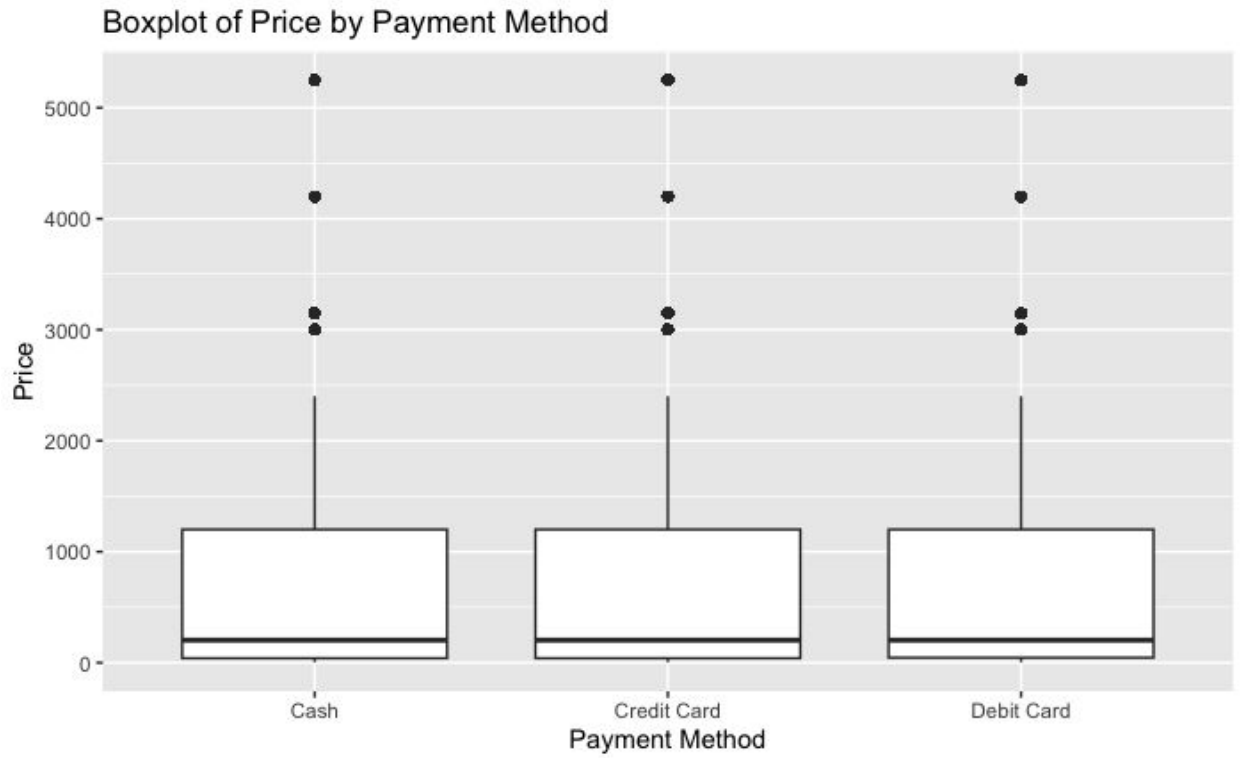
```
shopping_model2 <- lm(Price ~ Payment_Method, data = customerDF)
```

In this hypothesis, we examine whether the payment method used (Cash, Credit Card, or Debit Card) significantly impacts customers' purchase amount (Price). The regression model includes Payment_Method as the predictor variable and Price as the response variable.

Output 2

Hypothesis 2: Payment Method Affects Purchase Amount

```
shopping_model2 <- lm(price ~ payment_method, data = customerDF)
summary(shopping_model2)
## Call:
## lm(formula = price ~ payment_method, data = customerDF)

## Residuals:
##    Min     1Q Median     3Q    Max
## -685.6 -645.4 -483.7  509.5 4563.0

## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               690.823      4.464 154.743   <2e-16 ***
## payment_methodCredit Card  -2.281      6.730  -0.339    0.735
## payment_methodDebit Card   -3.794      8.003  -0.474    0.635
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 941.2 on 99454 degrees of freedom
## Multiple R-squared:  2.572e-06,  Adjusted R-squared:  -1.754e-05
## F-statistic: 0.1279 on 2 and 99454 DF,  p-value: 0.88
```



Based on these results, we can conclude that Payment Method (Credit Card or Debit Card) does not significantly impact Purchase Amount in this dataset, as it is not a statistically significant predictor.

Hypothesis 3 - Shopping Mall Location Influences Purchase Frequency

Null Hypothesis (H0): Shopping mall location does not significantly influence purchase frequency.

Alternative Hypothesis (H1): Shopping mall location significantly influences purchase frequency.

Regression model 3:

```
shopping_model3 <- lm(Quantity ~ Shopping_Mall, data = customerDF)
```

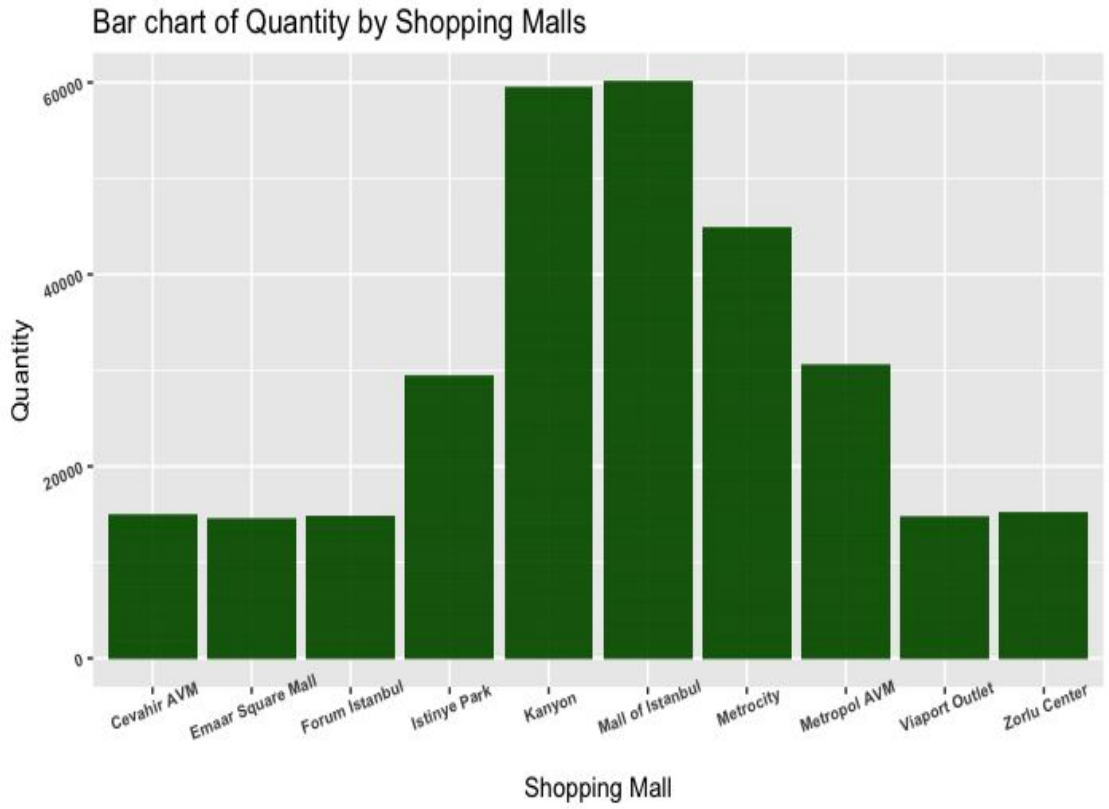
In this hypothesis, we are investigating whether the location of the shopping mall (where the transaction was made) significantly impacts customers' purchase frequency (Quantity). The regression model includes Shopping_Mall as the predictor variable and Quantity as the response variable.

Output 3

Hypothesis 3: Shopping Mall Location Influences Purchase Frequency

```
shopping_model3 <- lm(quantity ~ shopping_mall, data = customerDF)
summary(shopping_model3)
```

```
##
## Call:
## lm(formula = quantity ~ shopping_mall, data = customerDF)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.01429 -1.01247 -0.00177  1.00481  2.00926
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   2.9951913  0.0200017 149.747   <2e-16 ***
## shopping_mallEmaar Square Mall  0.0189429  0.0285501   0.663    0.507
## shopping_mallForum Istanbul     0.0070322  0.0283496   0.248    0.804
## shopping_mallIstinye Park       0.0172818  0.0245808   0.703    0.482
## shopping_mallKanyon             0.0042033  0.0223785   0.188    0.851
## shopping_mallMall of Istanbul   0.0190994  0.0223650   0.854    0.393
## shopping_mallMetrocity         -0.0044512  0.0230887  -0.193    0.847
## shopping_mallMetropol AVM       0.0094342  0.0244250   0.386    0.699
## shopping_mallViaport Outlet    -0.0004823  0.0283973  -0.017    0.986
## shopping_mallZorlu Center       0.0065821  0.0281694   0.234    0.815
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.413 on 99447 degrees of freedom
## Multiple R-squared:  3.622e-05,  Adjusted R-squared:  -5.428e-05
## F-statistic: 0.4002 on 9 and 99447 DF,  p-value: 0.9356
```



Based on the above analysis we can conclude that the regression result indicates that there is no statistically significant relationship between the shopping mall location and the quantity purchased. However, a bar plot shows that Mall of Istanbul and Mall of Kanyon have the highest purchase quantity based on the data in the dataset. This suggests that a bar plot only shows the distribution of the data and does not necessarily provide evidence for causality or relationships between variables.

Hypothesis 4 - Product Category Affects Purchase Amount

Null Hypothesis (H0): Product category does not significantly affect the purchase.

Alternative Hypothesis (H1): Product category significantly affects the purchase amount.

Regression model 4:

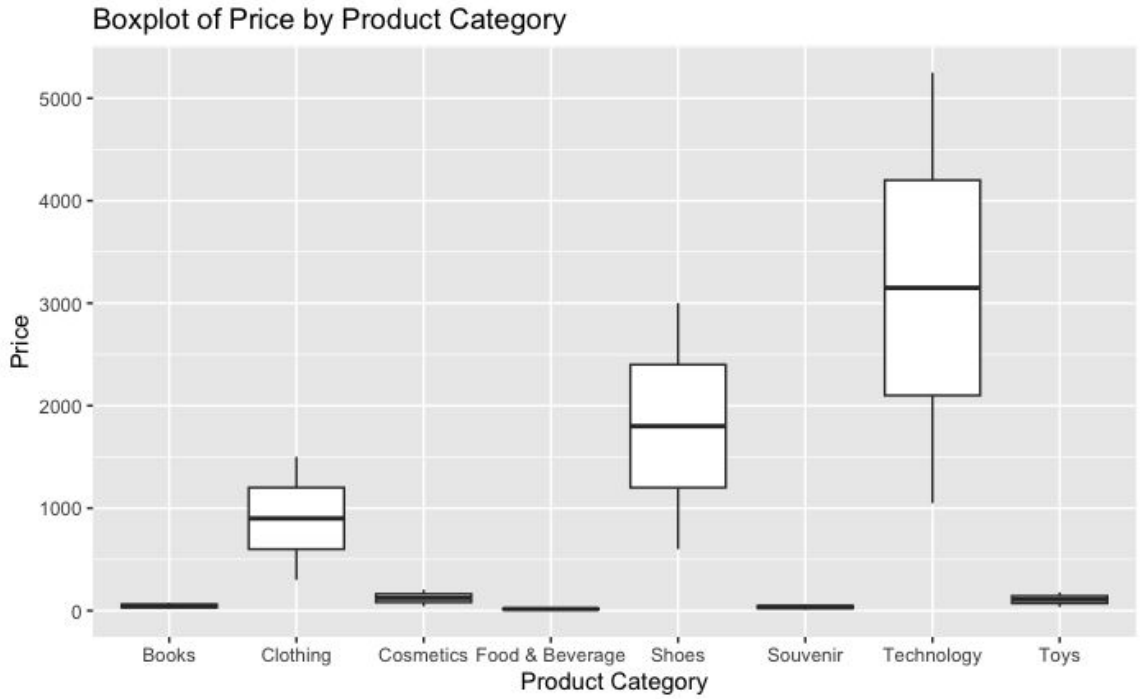
```
shopping_model4 <- lm(Price ~ Category, data = customerDF)
```

In this hypothesis, we are exploring whether the product category (e.g., clothing, electronics, groceries, etc.) significantly impacts customers' purchase amount (Price). The regression model includes Category as the predictor variable and Price as the response variable.

Output 4

Hypothesis 4: Product Category Affects Purchase Amount

```
shopping_model4 <- lm(price ~ category, data = customerDF)
summary(shopping_model4)

##
## Call:
## lm(formula = price ~ category, data = customerDF)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2106.94   -41.13    -0.47    40.19  2093.06
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                45.569      7.021   6.491 8.59e-11 ***
## categoryClothing          855.515      7.511 113.908  < 2e-16 ***
## categoryCosmetics          76.880      8.096   9.496  < 2e-16 ***
## categoryFood & Beverage   -29.897      8.118  -3.683 0.000231 ***
## categoryShoes            1761.820      8.588 205.142  < 2e-16 ***
## categorySouvenir          -10.674      9.920  -1.076 0.281905
## categoryTechnology       3111.367      9.921 313.604  < 2e-16 ***
## categoryToys               62.165      8.581   7.245 4.37e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 495.5 on 99449 degrees of freedom
## Multiple R-squared:  0.7229, Adjusted R-squared:  0.7228
## F-statistic: 3.706e+04 on 7 and 99449 DF,  p-value: < 2.2e-16
```



In conclusion, the results suggest that product category significantly affects purchase amount, with technology products having the highest impact, followed by shoes and clothing. At the same time, food & beverage may hurt the purchase amount. Based on the results, the souvenir category may not significantly affect the purchase amount.

H. Cluster Analysis

Cluster Analysis

We are doing a k mean clustering analysis on our data set containing customer demographic and purchase behavior data.

The resulting clusters can be used to identify customer segments with different preferences and behavior patterns. The scatter plot matrix can be used to visually inspect the relationships between the variables and how they relate to the identified clusters.

I. Coordinating MySQL & R

R Analysis on SQL Database

- Step 1: Import the SQL (Part A) database into R

```
myCon<- dbConnect (MySQL(),user = "s23g_dl1123_dbu", password = "8Cq3Lu9O",
                   host = "datar-t.cbjtth9ysjtl.us-east-2.rds.amazonaws.com",
                   dbname = "s23g_Team05_db")
dbListTables(myCon)
```

```
## [1] "Category"    "Channel"     "Clothgender" "Customer"    "Payment"
## [6] "Product"     "Season"      "Transaction"
```

- Step 2: Import data from 4 tables into 1 data frame in R

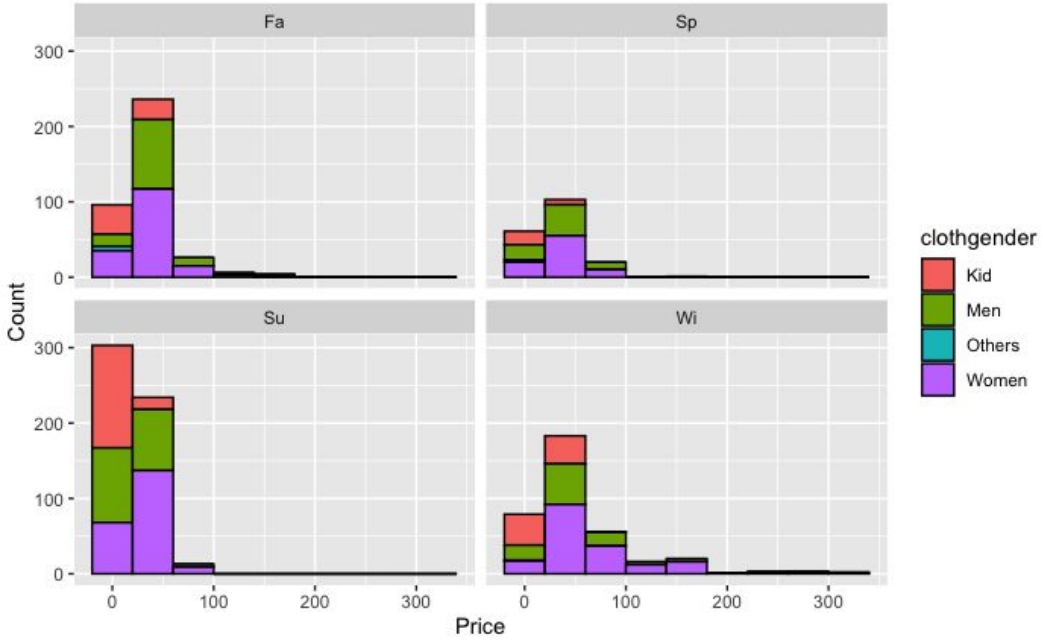
```
Query<- "SELECT Product.*, Season.season, category_name, clothgender
FROM Product INNER JOIN Season USING (ssn_id)
INNER JOIN Category USING (category_id)
INNER JOIN Clothgender USING (clothgender_id);"

ProductDB <- dbGetQuery(conn = myCon, statement = Query)
```

Analysis through 4 tables

Perform analysis by showing head of database and histograms

```
head(ProductDB)
```

```
##   product_id                     product_name price category_id ssn_id
## 1     419495                    MD Core T 29K  9.45           4      2
## 2     420235                  MD W's belt2990 39.95           6      1
## 3     422267                    W's Inner 79K  9.98           7      2
## 4     422813                  W's bottoms 199K 27.57           4      2
## 5     422962          MD DRY-EX S/S polo shirt 24.95           4      2
## 6     422989 SUPIMA cotton V neck S/S T-shirt 12.45           4      2
##   clothgender_id season  category_name clothgender
## 1              3     Su      cut & sewn         Men
## 2              2     Sp     accessories       Women
## 3              2     Su inner & living       Women
## 4              2     Su      cut & sewn       Women
## 5              3     Su      cut & sewn         Men
## 6              3     Su      cut & sewn         Men
```

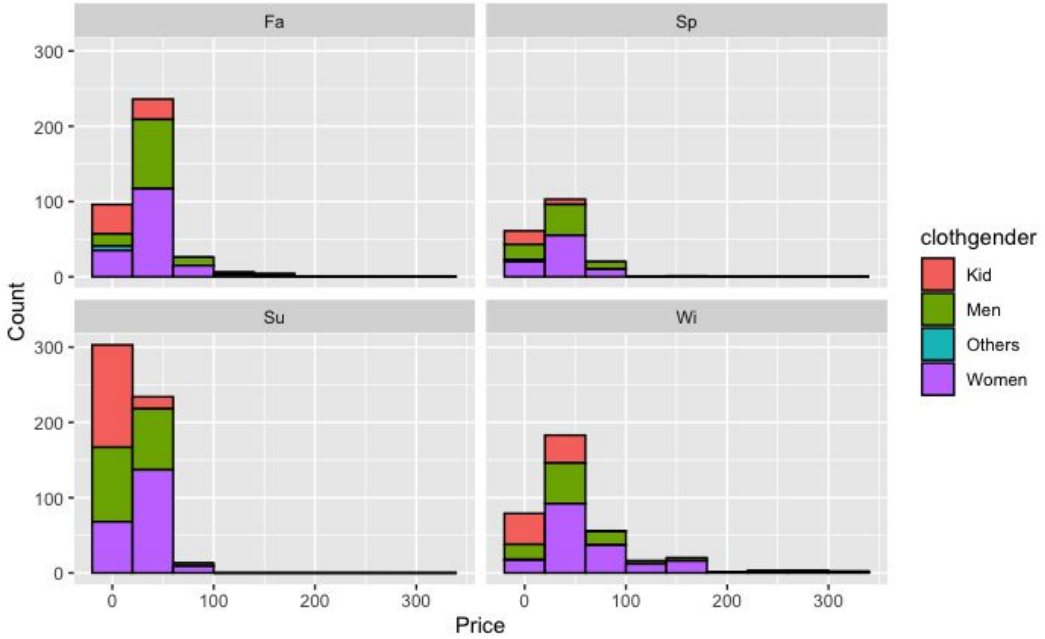



```
ggplot(ProductDB, aes(x = price, fill = clothgender)) +
geom_histogram(binwidth = 40, color= "black") + labs(x =
"Price", y = "Count") +  facet_wrap(~season)
```

Analysis through 4 tables

Output Analysis:

- Maximum shopping happens during Summer and then Fall.
- Price of Winter clothes are distributed well- items of higher prices are also being sold. This may be because winter wears like jackets and boots generally fall in the higher range.
- In Summer and Fall, most items sold are in the price range of $0-100.
- Most minor sales happen in Spring.
- Shopping for kids takes place in Summer the most.



Analysis through from Transaction and Payment

Perform analysis by showing head of database from table Transaction and table Payment

```
Query2 <- "SELECT Transaction.*, Payment.*
 FROM Transaction INNER JOIN Payment USING (payment_id);"


TransDB<- dbGetQuery(conn = myCon, statement = Query2)

head(TransDB)

##   transaction_id    transaction_date customer_id product_id quantity
## 1              1 2022-05-01 00:00:00         916     442479       10
## 2              2 2022-12-14 00:00:00         171     448874        9
## 3              3 2022-09-14 00:00:00         951     448874        2
## 4              4 2022-04-08 00:00:00         158     449824       10
## 5              5 2022-09-30 00:00:00         374     448874        5
## 6              6 2022-03-13 00:00:00         569     448874        4
##   total_amount channel_id payment_id payment_id payment_type
## 1       399.50          1          3          3   Mastercard
## 2       224.55          3          5          5  Credit Card
## 3        49.90          2          3          3   Mastercard
## 4       399.50          2          4          4       Coupon
## 5       124.75          3          5          5  Credit Card
## 6        99.80          1          2          2         VISA
```

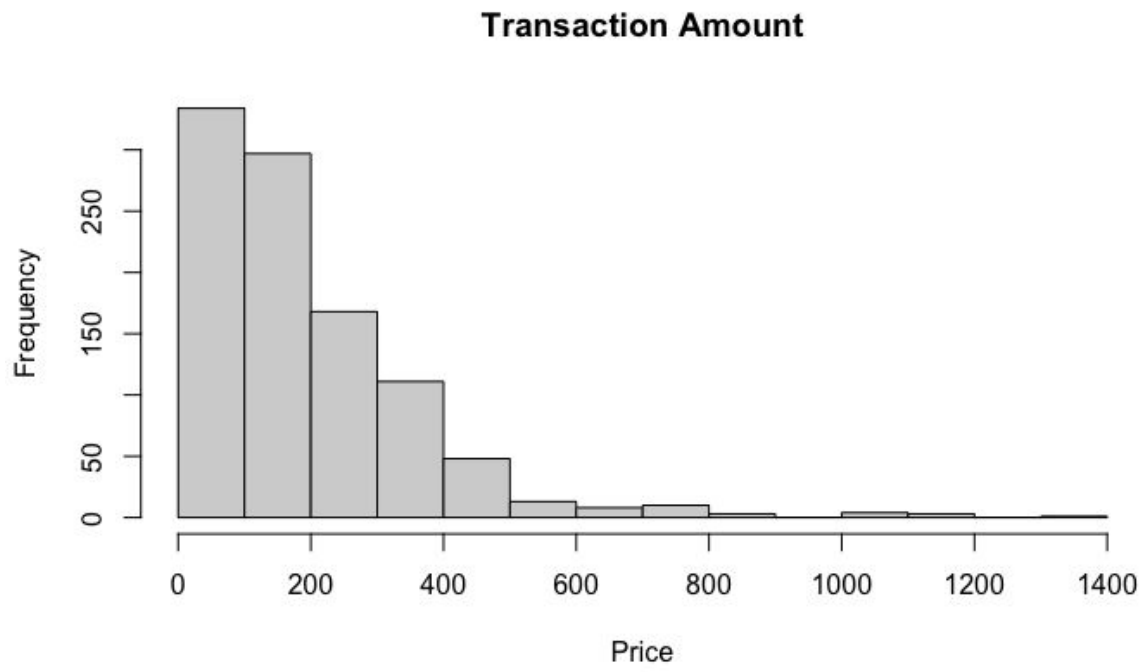
Analysis through from Transaction and Payment

1) Analysis of the total amount distributed:

The histogram shows that maximum transactions occur in the price range of 0-200. And there are moderate transactions in the field of $200-500.

But above $500, there are very few transactions.

```
hist(TransDB$total_amount, main = "Transaction
Amount", xlab = "Price")
```



Analysis through from Transaction and Payment

2) Analyze a correlation between transaction amount and channel id:

So, we see that R square is very low; hence it could be a better fit. Also, the p-value is high. Hence it is not statistically significant.

The billing amount is not dependent on channel type.

```
try1Lm <- lm(total_amount ~ channel_id, TransDB)
coefficients(try1Lm)

## (Intercept)  channel_id
##  205.960517   -1.910776
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  205.961     14.555  14.150   <2e-16 ***
## channel_id    -1.911      6.646  -0.288    0.774
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 171.2 on 998 degrees of freedom
## Multiple R-squared:  8.283e-05,  Adjusted R-squared:  -0.0009191
## F-statistic: 0.08267 on 1 and 998 DF,  p-value: 0.7738
```

J. Importing HTML into R

Importing HTML into R

- We have chosen an HTML table that lists the largest retail companies in the world by revenue.

```
url8<- "https://en.wikipedia.org/wiki/List_of_largest_retail_companies"
tabs1<-getURL(url8)
RetailCo<- readHTMLTable(tabs1, which = 1, header = FALSE, stringsAsFactors = FALSE)
head(RetailCo)
```

```
##     V1             V2                                V3
## 1 Rank           Name        Dominant operational format
## 2    1        Walmart Hypermarket/Supercenter/Superstore
## 3    2         Amazon               Non-Store,E-commerce
## 4    3         Costco        Cash & Carry/Warehouse Club
## 5    4 Schwarz Gruppe                    Discount Store
## 6    5 The Home Depot                  Home Improvement
##                                V4               V5          V6            V7
## 1 Retail revenue (US$ millions) Net profit margin Headquarters          <NA>
## 2                559,151              2.5%  Bentonville  United States
## 3                213,573              5.5%      Seattle  United States
## 4                166,761              2.4%     Issaquah  United States
## 5                144,254               ...   Neckarsulm        Germany
## 6                132,110              9.7%      Atlanta  United States
```

Business Conclusion

Lessons Learned:

1. R is a **powerful tool** for visualizing data frames and helps immensely in decision-making and drawing insights.
2. Presenting data and getting useful insights is **simple, fast, and efficient**.
3. It can help identify potential Patterns, Associations, and Clusters, which otherwise are very hard to locate.
4. Descriptive statistics may show the relationship between variables; however, in doing regression, the relationship between dependent and independent variables may not be statistically significant.

Challenges:

1. Data must be structured and accurate to derive conclusions. With meaningful data, we can avoid issues in our further analysis.
2. We must handle categorical and nominal data well before analyzing. Sometimes, we have numerical data (ranks, etc), but their regression doesn't make sense unless we convert them into more meaningful nomenclature.
3. While doing analysis, sometimes there arises a need for more data to help the analysis. Hence, a proper Business Understanding is very important before we start the process of data collation and analysis.

Thank You!