Case Study - Data Exploration

For the analysis part, we will string out the most important components of our data to answer our business objectives.

Let’s load our data into SQL and check the first 5 rows to make sure it imported well.

```
SELECT * FROM superstore LIMIT 5;
```

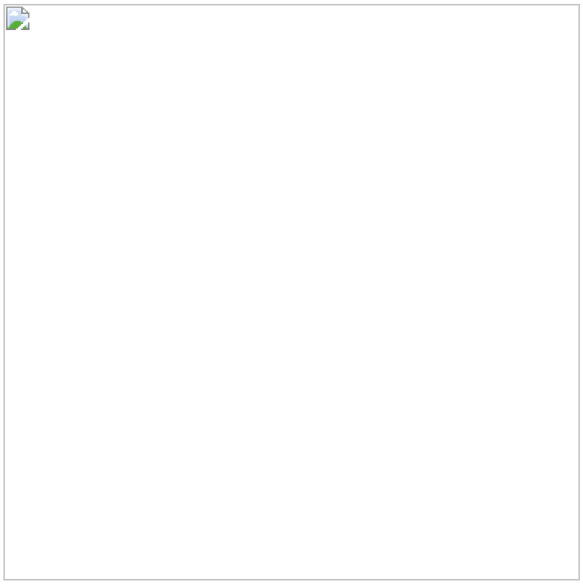
Okay, let’s perform an exploratory data analysis with our input on the superstore dataset. A list of tasks will be answered followed by the query input and query result.

1. **What are total sales and total profits of each year?**

The years were grouped by order date, so we can observe data for the year 2014, 2015, 2016 and 2017.

```
SELECT DATE_TRUNC('year', orderdate) AS year,
SUM(sales) AS total_sales,
SUM(profit) AS total_profit
FROM superstore
GROUP BY year
ORDER BY year ASC;
```

This query produced the following result:

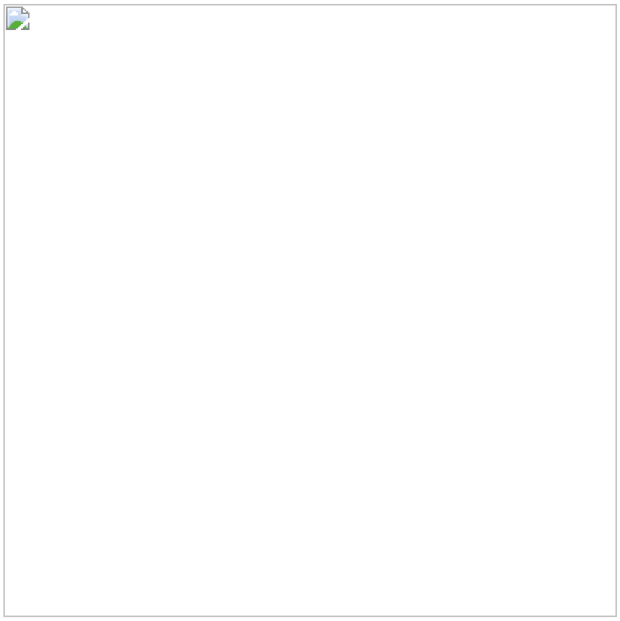


The data above shows how the profits over the years have steadily increased with each year being more profitable than the other despite having a fall in sales in 2015, our financial performance

2. What are the total profits and total sales per quarter?

This is done to see the periods where our company has been the most impactful.

```
SELECT
date_part('year', orderdate) AS year,
CASE
WHEN date_part('month', orderdate) IN (1,2,3) THEN 'Q1'
WHEN date_part('month', orderdate) IN (4,5,6) THEN 'Q2'
WHEN date_part('month', orderdate) IN (7,8,9) THEN 'Q3'
ELSE 'Q4'
END AS quarter,
SUM(sales) AS total_sales,
SUM(profit) AS total_profit
FROM superstore
GROUP BY year, quarter
ORDER BY year, quarter;
```

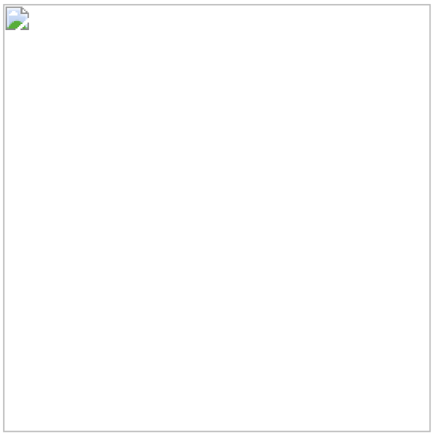


3. What region generates the highest sales and profits ?

```
SELECT region, SUM(sales) AS total_sales, SUM(profit) AS total_profits
FROM superstore
GROUP BY region
ORDER BY total_profits DESC;
```

Let’s observe each regions profit margins for further analysis with the following code:

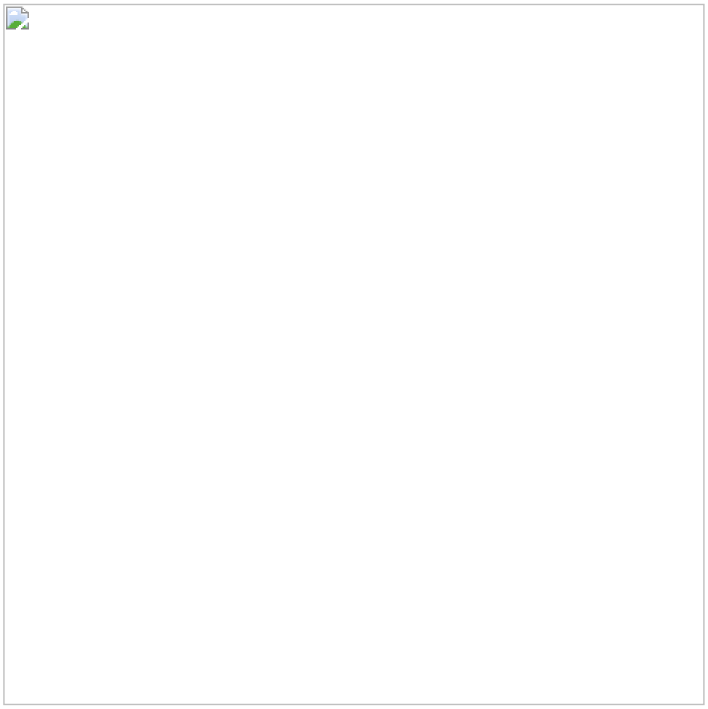
```
SELECT region, ROUND((SUM(profit) / SUM(sales)) * 100, 2) as profit_margin
FROM superstore
GROUP BY region
ORDER BY profit_margin DESC
```



4. What state and city brings in the highest sales and profits ?

States

```
SELECT State, SUM(Sales) as Total_Sales, SUM(Profit) as Total_Profits, ROUND((SUM(profit) / SUM(sales)) * 100, 2) as profit_margin
FROM superstore
GROUP BY State
ORDER BY Total_Profits DESC
LIMIT 10;
```



Let’s observe our bottom 10 States:

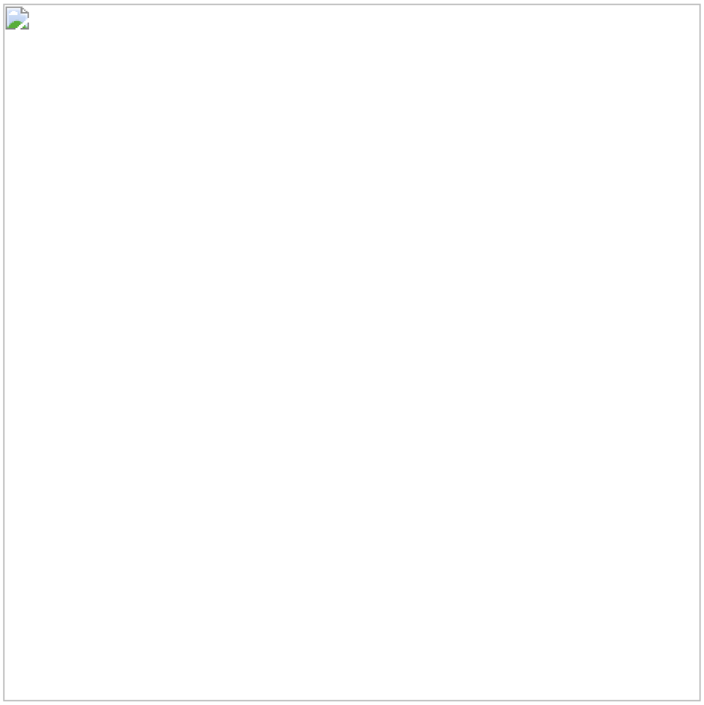
```
SELECT State, SUM(Sales) as Total_Sales, SUM(Profit) as Total_Profits
FROM superstore
GROUP BY State
```

```
ORDER BY Total_Profits ASC
LIMIT 10;
```

Cities

The top cities are found with the code below:

```
SELECT City, SUM(Sales) as Total_Sales, SUM(Profit) as Total_Profits, ROUND((SUM(profit) / SUM(sales)) * 100, 2) as profit_margin
FROM superstore
GROUP BY City
ORDER BY Total_Profits DESC
LIMIT 10;
```



The top 3 cities that we should focus on are New York City, Los Angeles and Seattle.

The bottom 10 cities are:

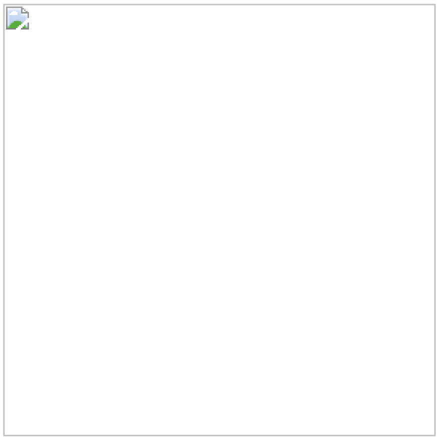
```
SELECT City, SUM(Sales) as Total_Sales, SUM(Profit) as Total_Profits
FROM superstore
GROUP BY City
ORDER BY Total_Profits ASC
LIMIT 10;
```

5. The relationship between discount and sales and the total discount per category

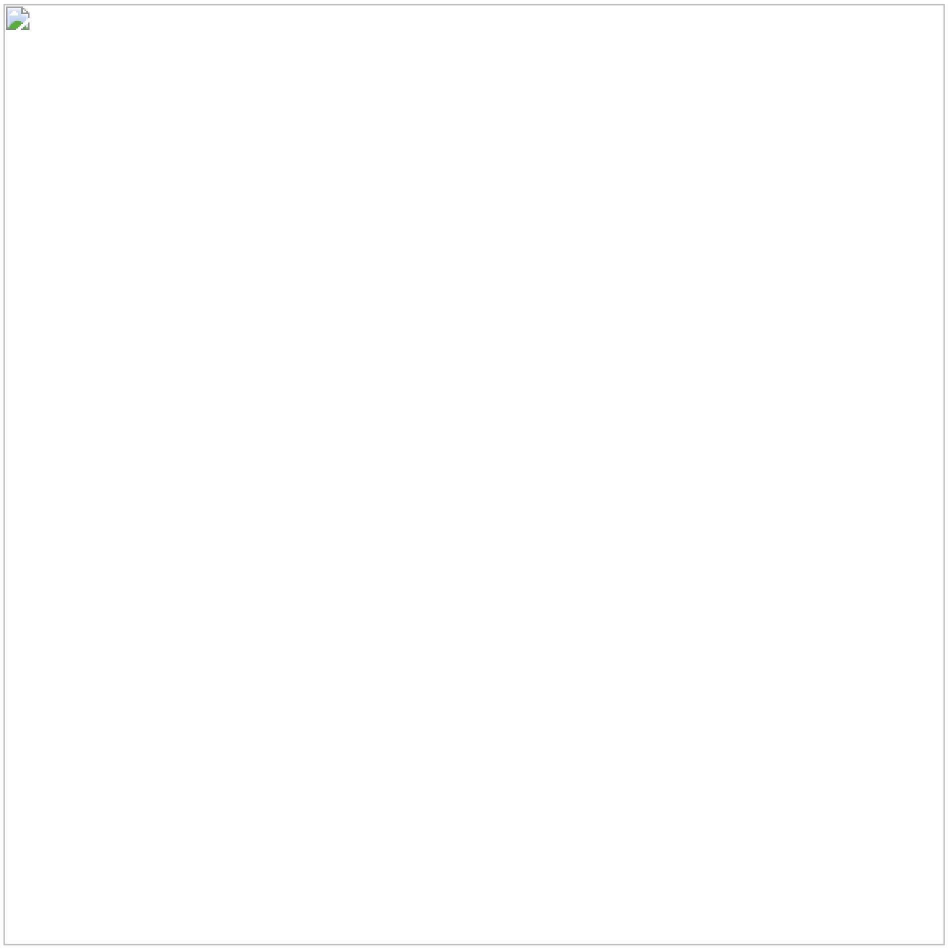
First, let’s observe the correlation between discount and average sales to understand how impactful one is to the other.

```
SELECT Discount, AVG(Sales) AS Avg_Sales
FROM superstore
GROUP BY Discount
ORDER BY Discount;
```

This produces the following:



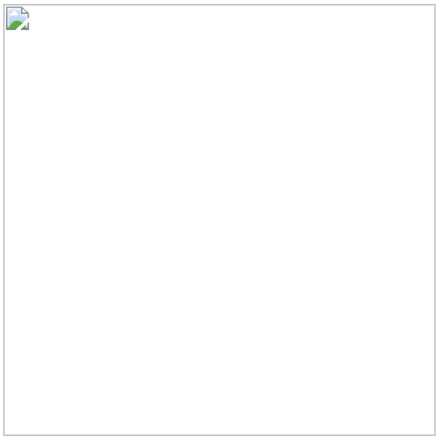
Seems that for each discount point, the average sales seem to vary a lot. Let’s check the correlation with a graph in Excel.



They almost have no linear relationship. This noted by the correlation coefficient of -0.3 and the shape of the graph. However we can at least observe that at a 50% discount, (0.5 * 100 to convert it to percentage) our average sales are the highest it can be. Maybe it is a psychology technique or it’s just the right product category that is discounted.

Let’s observe the total discount per product category:

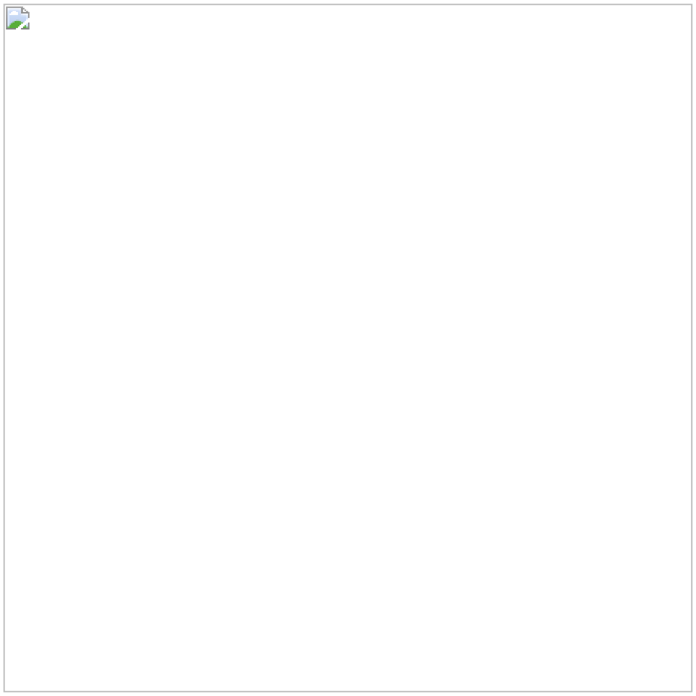
```
SELECT category, SUM(discount) AS total_discount
FROM superstore
GROUP BY category
ORDER BY total_discount DESC;
```



6. What category generates the highest sales and profits in each region and state ?

First, let’s observe the total sales and total profits of each category with their profit margins:

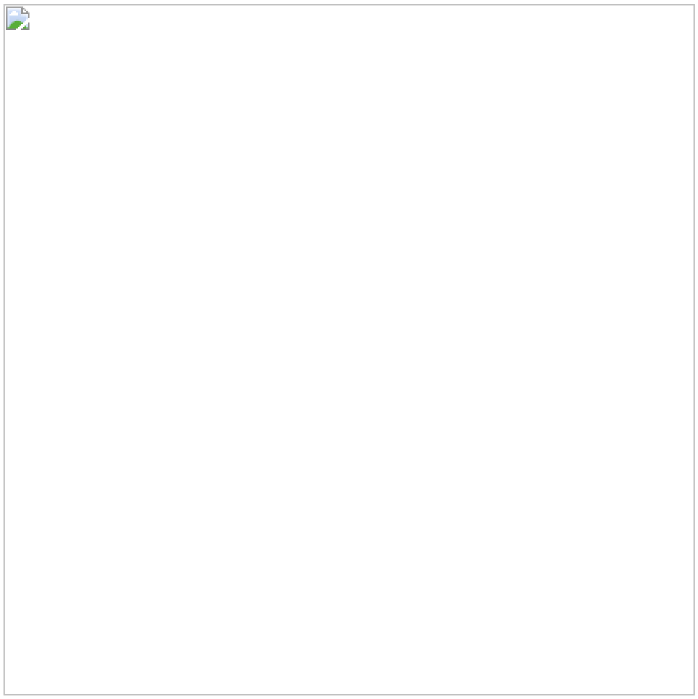
```
SELECT category, SUM(sales) AS total_sales, SUM(profit) AS total_profit, ROUND(SUM(profit)/SUM(sales)*100, 2) AS profit_margin
FROM superstore
GROUP BY category
ORDER BY total_profit DESC;
```



7. What subcategory generates the highest sales and profits in each region and state ?

Let’s observe the total sales and total profits of each subcategory with their profit margins:

```
SELECT subcategory, SUM(sales) AS total_sales, SUM(profit) AS total_profit, ROUND(SUM(profit)/SUM(sales)*100, 2) AS profit_margin
FROM superstore
GROUP BY subcategory
ORDER BY total_profit DESC;
```



8. What are the names of the products that are the most and least profitable to us?

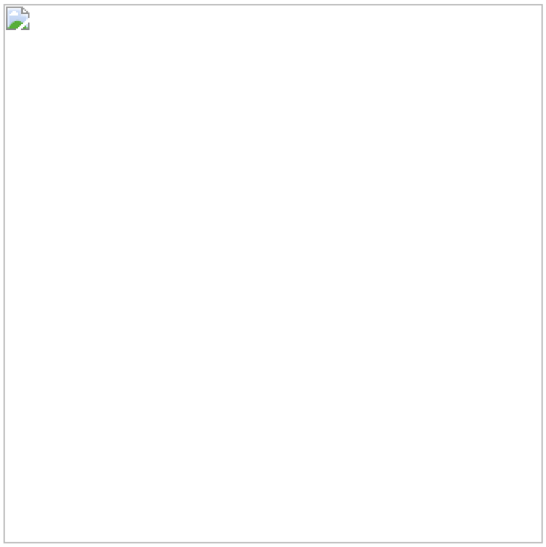
Let’s verify this information:

```
SELECT productname, SUM(sales) AS total_sales, SUM(profit) AS total_profit
FROM superstore
GROUP BY productname
ORDER BY total_profit DESC;
```

9. What segment makes the most of our profits and sales ?

This can be verified with the help of the following query:

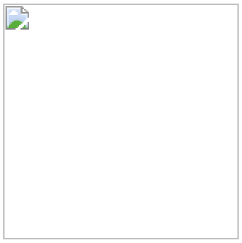
```
SELECT segment, SUM(sales) AS total_sales, SUM(profit) AS total_profit
FROM superstore
GROUP BY segment
ORDER BY total_profit DESC;
```



10. How many customers do we have (unique customer IDs) in total and how much per region and state?

This can be solved with the following;

```
SELECT COUNT(DISTINCT customerid) AS total_customers
FROM superstore;
```

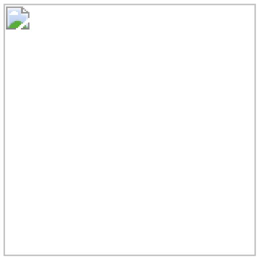


We’ve had 793 customers between 2014 and 2017. Regionally, we had the following:

12. Average shipping time per class and in total

Finally, the average shipping time regardless of the shipping mode that is chosen is found with the following function:

```
SELECT ROUND(AVG(shipdate - orderdate),1) AS avg_shipping_time
FROM superstore
```



The shipping time in each shipping mode is:

```
SELECT shipmode,ROUND(AVG(shipdate - orderdate),1) AS avg_shipping_time
FROM superstore
GROUP BY shipmode
ORDER BY avg_shipping_time
```

Finally, for our clientele, we have 793 customers total, and we have the most customers in California, New York and Texas. The case of Texas is pretty ironic since it is also the state that losses us the most money.

So we must take a critical decision about Texas first as we absolutely can’t break through now. California and New York are pretty obvious, we have to be outstanding and be the best of what there is to offer in our respective niche.

Thank you for your time!.