Case Study - Data Exploration

Created	@September 19, 2023 11:30 AM
Tags	

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;
```

	rowid integer	orderid character varying (20)	orderdate date	shipdate date	shipmode character varying (20)	customerid character varying (20)	customername character varying (100)
1	1	CA-2016-152156	2016-11-08	2016-11-11	Second Class	CG-12520	Claire Gute
2	2	CA-2016-152156	2016-11-08	2016-11-11	Second Class	CG-12520	Claire Gute
3	3	CA-2016-138688	2016-06-12	2016-06-16	Second Class	DV-13045	Darrin Van Huff
4	4	US-2015-108966	2015-10-11	2015-10-18	Standard Class	SO-20335	Sean O'Donnell
5	5	US-2015-108966	2015-10-11	2015-10-18	Standard Class	SO-20335	Sean O'Donnell

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:

	year timestamp with time zone	total_sales numeric	total_profit numeric
1	2014-01-01 00:00:00-06	484247.56	49544.06
2	2015-01-01 00:00:00-06	470532.46	61618.69
3	2016-01-01 00:00:00-06	609205.86	81795.27
4	2017-01-01 00:00:00-06	733215.19	93439.77

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;
```

	year double precision	quarter text	total_sales numeric	total_profit numeric
1	2014	Q1	74447.86	3811.20
2	2014	Q2	86538.77	11204.16
3	2014	Q3	143633.18	12804.73
4	2014	Q4	179627.75	21723.97
5	2015	Q1	68851.74	9264.94
6	2015	Q2	89124.28	12190.92
7	2015	Q3	130259.51	16853.70
8	2015	Q4	182296.93	23309.13
9	2016	Q1	93237.20	11441.39
10	2016	Q2	136082.33	16390.30
11	2016	Q3	143787.43	15823.63
12	2016	Q4	236098.90	38139.95
13	2017	Q1	123144.84	23506.21
14	2017	Q2	133764.33	15499.42
15	2017	Q3	196251.94	26985.26
16	2017	Q4	280054.08	27448.88

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;
```

	region character varying (20)	total_sales numeric	total_profits numeric
1	West	725457.93	108418.79
2	East	678781.36	91522.84
3	South	391721.90	46749.71
4	Central	501239.88	39706.45

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
```

	region character varying (20)	profit_margin numeric
1	West	14.94
2	East	13.48
3	South	11.93
4	Central	7.92

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;
```

	state character varying (20)	total_sales numeric	total_profits numeric	profit_margin numeric
1	California	457687.68	76381.60	16.69
2	New York	310876.20	74038.64	23.82
3	Washington	138641.29	33402.70	24.09
4	Michigan	76269.61	24463.15	32.07
5	Virginia	70636.72	18598.00	26.33
6	Indiana	53555.36	18382.97	34.33
7	Georgia	49095.84	16250.08	33.10
8	Kentucky	36591.75	11199.70	30.61
9	Minnesota	29863.15	10823.22	36.24
10	Delaware	27451.07	9977.37	36.35

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;
```

	state character varying (20)	total_sales numeric	total_profits numeric
1	Texas	170187.98	-25729.29
2	Ohio	78258.21	-16971.37
3	Pennsylvania	116512.02	-15560.04
4	Illinois	80166.16	-12607.89
5	North Carolina	55603.09	-7490.81
6	Colorado	32108.12	-6527.86
7	Tennessee	30661.92	-5341.66
8	Arizona	35282.02	-3427.87
9	Florida	89473.73	-3399.25
10	Oregon	17431.14	-1190.48

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;
```

	city character varying (50)	total_sales numeric	total_profits numeric	profit_margin numeric
1	New York City	256368.12	62037.08	24.20
2	Los Angeles	175851.33	30440.94	17.31
3	Seattle	119540.74	29156.13	24.39
4	San Francisco	112669.09	17507.39	15.54
5	Detroit	42446.94	13181.79	31.05
6	Lafayette	25036.20	10018.38	40.02
7	Jackson	24963.85	7581.67	30.37
8	Atlanta	17197.84	6993.69	40.67
9	Minneapolis	16870.54	6824.61	40.45
10	San Diego	47521.05	6377.24	13.42

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;
```

	city character varying (50)	total_sales numeric	total_profits numeric
1	Philadelphia	109077.09	-13837.83
2	Houston	64504.71	-10153.48
3	San Antonio	21843.54	-7299.06
4	Lancaster	9891.48	-7239.08
5	Chicago	48539.59	-6654.55
6	Burlington	21668.08	-3622.88
7	Dallas	20131.90	-2846.55
8	Phoenix	11000.27	-2790.85
9	Aurora	11656.47	-2691.76
10	Jacksonville	44713.18	-2323.80

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
```

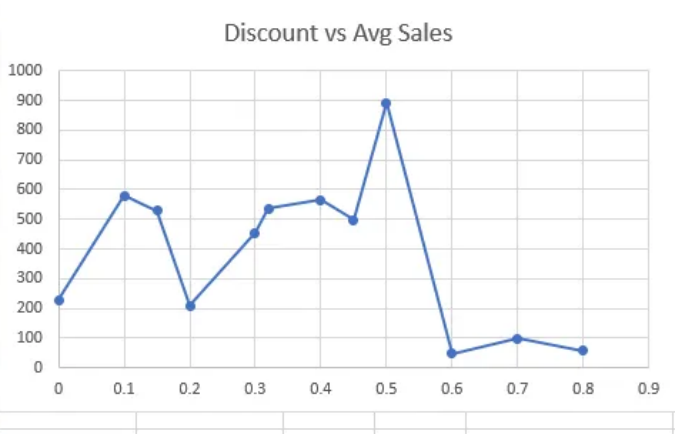
```
GROUP BY Discount
ORDER BY Discount;
```

This produces the following:

	discount numeric (4,2)	avg_sales numeric
1	0.00	226.7420737807419758
2	0.10	578.3968085106382979
3	0.15	529.9728846153846154
4	0.20	209.0769155045118950
5	0.30	454.7434361233480176
6	0.32	536.7944444444444444
7	0.40	565.1350970873786408
8	0.45	498.6345454545454545
9	0.50	892.7068181818181818
10	0.60	48.1498550724637681
11	0.70	97.1779904306220096
12	0.80	56.5456000000000000

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

Discount	Avg_sales
0	226.7420738
0.1	578.3968085
0.15	529.9728846
0.2	209.0769155
0.3	454.7434361
0.32	536.7944444
0.4	565.1350971
0.45	498.6345455
0.5	892.7068182
0.6	48.14985507
0.7	97.17799043
0.8	56.5456
Correlation	-0.342620367



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;
```

	category character varying (50)	total_discount numeric
1	Office Supplies	947.80
2	Furniture	368.89
3	Technology	244.40

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;
```

	category character varying (50)	total_sales numeric	total_profit numeric	profit_margin numeric
1	Technology	836154.10	145455.66	17.40
2	Office Supplies	719046.99	122490.88	17.04
3	Furniture	741999.98	18451.25	2.49

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;
```

	subcategory character varying (50)	total_sales numeric	total_profit numeric	profit_margin numeric
1	Copiers	149528.01	55617.90	37.20
2	Phones	330007.10	44516.25	13.49
3	Accessories	167380.31	41936.78	25.05
4	Paper	78479.24	34053.34	43.39
5	Binders	203412.77	30221.64	14.86
6	Chairs	328449.13	26590.15	8.10
7	Storage	223843.59	21279.05	9.51
8	Appliances	107532.14	18138.07	16.87
9	Furnishings	91705.12	13059.25	14.24
10	Envelopes	16476.38	6964.10	42.27
11	Art	27118.80	6527.96	24.07
12	Labels	12486.30	5546.18	44.42
13	Machines	189238.68	3384.73	1.79
14	Fasteners	3024.25	949.53	31.40
15	Supplies	46673.52	-1188.99	-2.55
16	Bookcases	114880.05	-3472.56	-3.02
17	Tables	206965.68	-17725.59	-8.56

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;
```

	productname character varying (150)	total_sales numeric	total_profit numeric
1	Canon imageCLASS 2200 Advanced Copier	61599.83	25199.94
2	Fellowes PB500 Electric Punch Plastic Comb Binding Machine with Manual Bind	27453.38	7753.06
3	Hewlett Packard LaserJet 3310 Copier	18839.68	6983.89
4	Canon PC1060 Personal Laser Copier	11619.83	4570.94
5	HP Designjet T520 Inkjet Large Format Printer - 24" Color	18374.90	4094.98
6	Ativa V4110MDD Micro-Cut Shredder	7699.89	3772.95
7	3D Systems Cube Printer, 2nd Generation, Magenta	14299.89	3717.97
8	Plantronics Savi W720 Multi-Device Wireless Headset System	9367.29	3696.28
9	Ibico EPK-21 Electric Binding System	15875.92	3345.29
10	Zebra ZM400 Thermal Label Printer	6965.70	3343.53
11	Honeywell Enviracaire Portable HEPA Air Cleaner for 17' x 22' Room	11304.44	3247.02
12	Hewlett Packard 610 Color Digital Copier / Printer	8899.82	3124.94
13	Plantronics CS510 - Over-the-Head monaural Wireless Headset System	10822.36	3085.04
14	Canon Imageclass D680 Copier / Fax	8959.87	2799.97
15	Fellowes PB300 Plastic Comb Binding Machine	8070.20	2518.06

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;
```

	segment character varying (20)	total_sales numeric	total_profit numeric
1	Consumer	1161401.34	134119.33
2	Corporate	706146.44	91979.45
3	Home Office	429653.29	60299.01

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;
```

	total_customers bigint
1	793

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
```

	avg_shipping_time numeric
1	4.0

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
```

	shipmode character varying (20)	avg_shipping_time numeric
1	Same Day	0.0
2	First Class	2.2
3	Second Class	3.2
4	Standard Class	5.0

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!.