

DSO 552 Group Project (Due Thursday, 2/22 9:30 AM PST)

Group 3 Members: Syed Muhammad Kumail Abbas, Zeze Gu, Nida Iqbal, Alexandra Li, Jacqueline Min, Hyun-Bi Park, Nico Santoso, Luzhou Shen

Part 1:

1. How big is the customer base ParchandPosey (i.e.how many customers/ accounts does the company have?) (1 point)

`select count(distinct(id)) as customer_base from accounts`

	customer_base bigint
1	351

2. How Many Areas Do They Sell At? (1point)

`select count(*) as number_of_sales_areas from region`

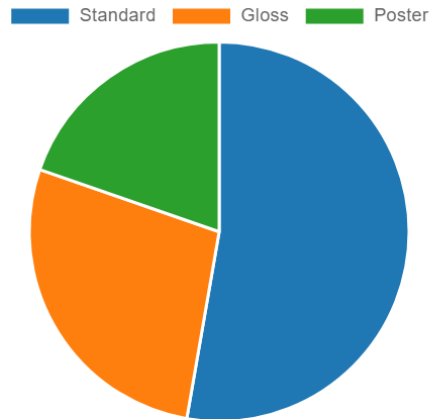
	number_of_sales_areas bigint
1	7

3. Look Into The Revenue Streams:

a. How many types of paper do they sell and what percentage each one of them makes out of the total quantity sold? Provide a visualization that illustrates the results (e.g. pie chart, bar plot, or any chart of your choice) (1.5 point)

`select sum(standard_qty) as standard_paper_quantity, sum(gloss_qty) as gloss_paper_quantity, sum.poster_qty) as poster_paper_quantity, sum(total) as total_paper_quantity, 100 * sum(standard_qty) / sum(total) as standard_paper_percentage, 100 * sum(gloss_qty) / sum(total) as gloss_paper_percentage, 100 * sum.poster_qty) / sum(total) as poster_paper_percentage from orders`

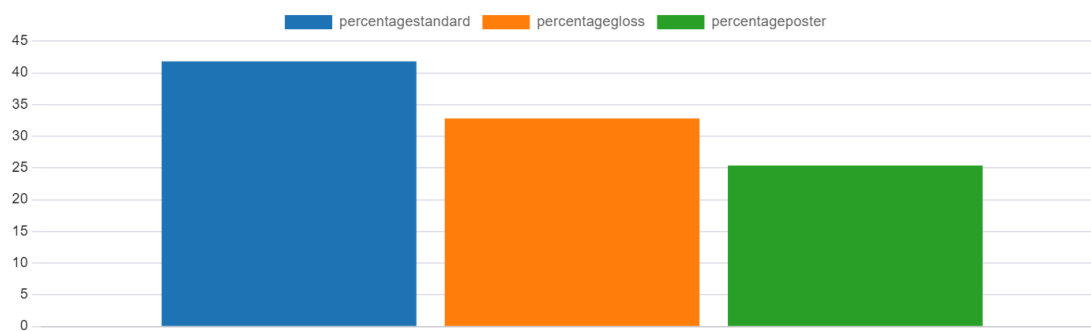
	standard_paper_quantity bigint	gloss_paper_quantity bigint	poster_paper_quantity bigint	total_paper_quantity bigint	standard_paper_percentage bigint	gloss_paper_percentage bigint	poster_paper_percentage bigint
1	1938346	1013773	723646	3675765	52	27	19



b. What percentage of revenues comes from which type of paper? Provide a visualization that illustrates the results (e.g. pie chart, bar plot, or any chart of your choice) (1.5 point)

```
select 100 * sum(standard_amt_usd) / sum(standard_amt_usd + gloss_amt_usd + poster_amt_usd) as standard_paper_revenue_percentage,
100 * sum(gloss_amt_usd) / sum(standard_amt_usd + gloss_amt_usd + poster_amt_usd) as gloss_paper_revenue_percentage,
100 * sum(poster_amt_usd) / sum(standard_amt_usd + gloss_amt_usd + poster_amt_usd) as poster_paper_revenue_percentage from orders
```

	standard_paper_revenue_percentage numeric	gloss_paper_revenue_percentage numeric	poster_paper_revenue_percentage numeric
1	41.7965196528821600	32.8118570030348791	25.3916233440829609



4. Is The Business Growing?

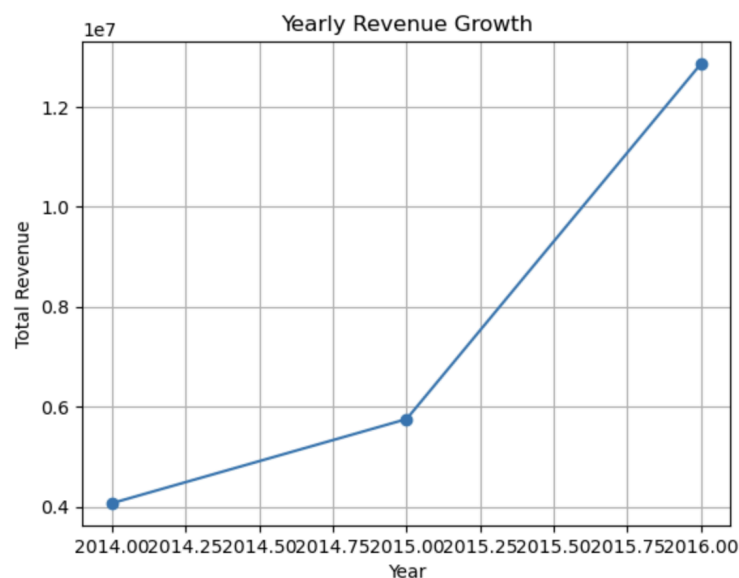
a. How have revenues been year over year? For this, only take into account years with full data (2017 just started, so we don't know how yearly revenues will be and 2013 seems to have data only from December). Provide a visualization that illustrates the results (e.g. line chart, bar plot, or any chart of your choice). (1.5 point)

```

select extract (year from occurred_at) as year, sum(total_amt_usd) as total_yearly_revenue
from orders
where extract (year from occurred_at) between 2014 and 2016
group by extract (year from occurred_at)

```

	year numeric 🔒	total_yearly_revenue numeric 🔒
1	2014	4069106.54
2	2015	5752004.94
3	2016	12864917.92



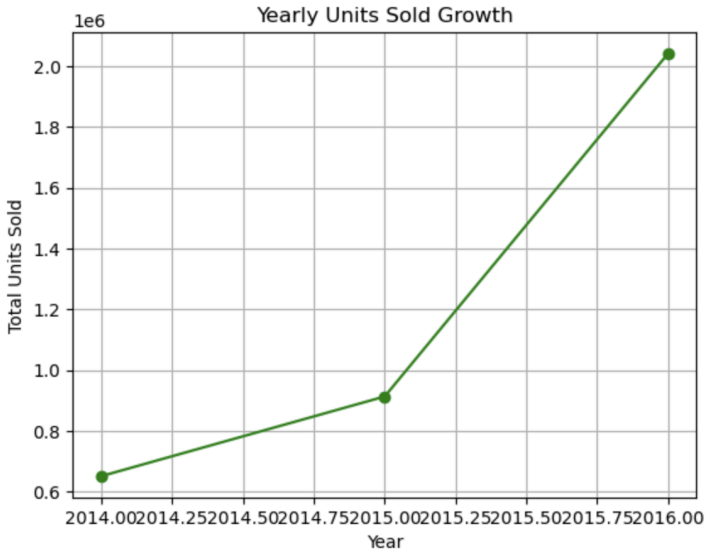
b. How have units sold evolved year over year? Here too, only take into account the past years' data. Provide a visualization that illustrates the results (e.g. line chart, bar plot, or any chart of your choice) (1.5 point)

```

select extract (year from occurred_at) as year, sum(total) as total_units_yearly_sold from orders
where extract (year from occurred_at) between 2014 and 2016
group by extract (year from occurred_at)

```

	year numeric 🔒	total_units_yearly_sold bigint 🔒
1	2014	650896
2	2015	912972
3	2016	2041600



5. How many sales reps do they have in each Region? Sort the result By Alphabetical order and include the regions that do not have any sales reps (1.5 point)

```
select region.name as region_name, count(sales_reps.name) as sale_reps from sales_reps
right join region on region.id = sales_reps.region_id
group by region.name
order by region.name
```

	region_name	sale_reps
	character	bigint
1	International	1
2	Midwest	9
3	North	0
4	Northeast	21
5	South	0
6	Southeast	10
7	West	10

6.

a. From Parch and Posey's leadership team you know that North, South and International are 3 newly added regions. If Dunder Mifflin decided to buy Parch and Posey, they would need to jump start sales in those areas. **How would you suggest reallocating sales reps from the old to the new regions to cover the needs of the latter**, i.e. which old regions would you recommend to pull sales reps from? To answer this question, compute in one query the following, only including data from the last year (year 2016):

- the total number of orders per region name
- the number of reps per region name
- the number of accounts per region name - the total revenues per region name
- the average revenues per region name (2 points)

```
select r.name, count(*) ordercount,
count(distinct(s.name)) repscount,
count(distinct(a.name)) accountcount,
sum(o.total_amt_usd) total_revenue,
avg(o.total_amt_usd) avg_revenue
from orders o
join accounts a on o.account_id=a.id
join sales_reps s on a.sales_rep_id = s.id
join region r on s.region_id = r.id
where extract(year from o.occurred_at) = 2016
group by r.name
```

	region_name character	total_orders bigint	number_of_reps bigint	number_of_accounts bigint	total_revenues numeric	average_revenues_per_region numeric
1	Midwest	483	9	41	1711747.25	3543.9901656314699793
2	Northeast	1196	21	97	3999036.82	3343.6762709030100334
3	Southeast	1110	10	86	3545487.49	3194.1328738738738739
4	West	968	10	93	3608646.36	3727.9404545454545455

b. Based on the previous result, compute also by region:

- average number of orders per representative (across all representatives) - average number of accounts handled per representative (across all representatives)
- average revenues per representative (across all representatives) (3 points)

	name character	ordercount bigint	repscount bigint	accountcount bigint	total_revenue numeric	avg_revenue numeric	avg_orders_per_rep bigint	avg_accounts_per_rep bigint	avg_revenue_per_rep numeric
1	Midwest	483	9	41	1711747.25	3543.9901656314699793	53	4	190194.138888888889
2	Northeast	1196	21	97	3999036.82	3343.6762709030100334	56	4	190430.324761904762
3	Southeast	1110	10	86	3545487.49	3194.1328738738738739	111	8	354548.749000000000
4	West	968	10	93	3608646.36	3727.9404545454545455	96	9	360864.636000000000

```
select name,
(ordercount/repscount) as avg_orders_per_rep,
(accountcount/repscount) as avg_accounts_per_rep,
(total_revenue/repscount) as avg_revenue_per_rep
from(
select r.name, count(*) ordercount,
count(distinct(s.name)) repscount,
count(distinct(a.name)) accountcount,
sum(o.total_amt_usd) total_revenue,
avg(o.total_amt_usd) avg_revenue
from orders o
```

```

join accounts a on o.account_id=a.id
join sales_reps s on a.sales_rep_id = s.id
join region r on s.region_id = r.id
where extract(year from o.occurred_at) = 2016
group by r.name)

```

c. Based on your calculations above, how would you recommend reallocating sales_reps to cover the new regions? (Any logical answer would work here, as long as it's backed up by the data and it makes sense) (1.5 points)

- **We recommend taking 11 sales reps from Northeast as well as 4 from Midwest to relocate them to the new regions**
- **Our initial suggestion will be 5 sales rep per new region until more data is available such as total number of accounts/orders in that region**
- **Based on the data from Southeast and West sales reps, we are confident that each sales reps can handle 8/9 accounts**

7. You suspect that accounts with the word group'attheendoftheirname are likely to bring in more revenues, since they may represent a group of multiple businesses. This would be useful to know, in order to try to understand if these accounts should be given more attention after a possible acquisition by Dunder Mifflin. To answer if this is true, create a new column in your output that is:



- 'group' if the name of the account ends with the word 'group'
- 'not group' otherwise

Then, based on the above result, compute the average (per account) revenues that came respectively from 'group' and from 'not group' accounts. (Hint: Here we would need 2 numbers, the average revenues for 'group' accounts and the average revenues for 'not group' accounts).

Finally, comment on the result and on whether your assumption was correct.
(2 points)

Need to discuss the per account vs over all orders but regardless the main takeaway is

Accounts with the word group at the end of their name had lower revenue. Accounts who do not have the word group at the end had higher revenues at [insert revenue]

	groupcheck text 	avg numeric 
1	not group	66351.025481927711
2	group	61831.742777777778

```

select groupcheck, avg(sum) from
(select account_id,groupcheck, sum(total_amt_usd) from
(select *,
case
    when a.name ilike '%group' or a.name ilike '%Group' then 'group'
    else 'not group'
end as groupcheck
from orders o
join accounts a on o.account_id=a.id)
group by account_id,groupcheck)
group by groupcheck

```

8. The Marketing team needs to focus on channels for the newly added sales regions, and because of its limited resources, it will have to deprioritize/deactivate temporarily some channels in the old areas. Specifically it decided to deactivate, for every old region, the channel that is used the least for web events in that region. Which channels should they deactivate in each region? Use a window function to give the answer here. (4 points)

Midwest - banner

Northeast - twitter

Southeast - twitter

West - banner

	region_name character	channel character	web_occur bigint	channel_rank bigint
1	Midwest	banner	59	1
2	Northeast	twitter	154	1
3	Southeast	twitter	127	1
4	West	banner	116	1

```

select * from (
SELECT
    r.name AS region_name,
    w.channel,
    COUNT(*) AS web_occur,
    RANK() OVER (PARTITION BY r.name ORDER BY COUNT(*) ASC) AS channel_rank
FROM web_events w
JOIN accounts a ON w.account_id = a.id
JOIN sales_reps s ON a.sales_rep_id = s.id
JOIN region r ON s.region_id = r.id
--WHERE r.name NOT IN ('North', 'South', 'International')
GROUP BY r.name, w.channel)
where channel_rank = 1

```