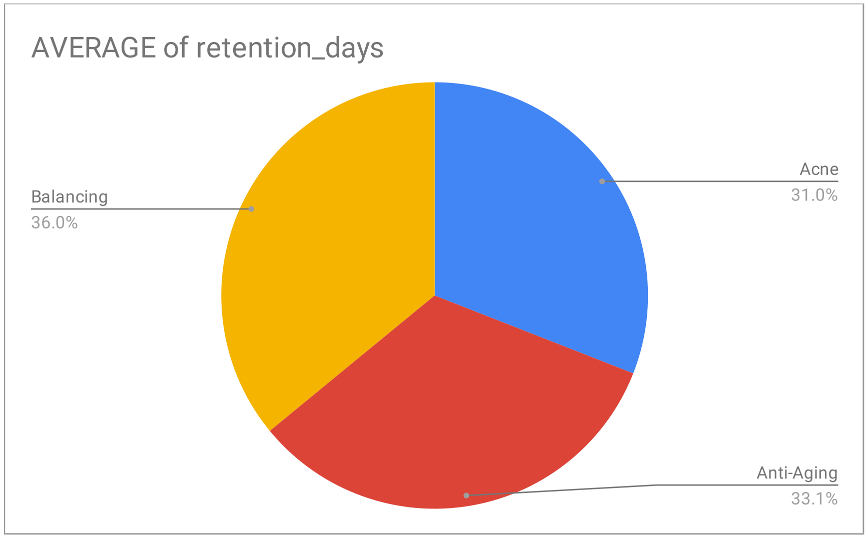
Assuming PostgreSQL used to query:

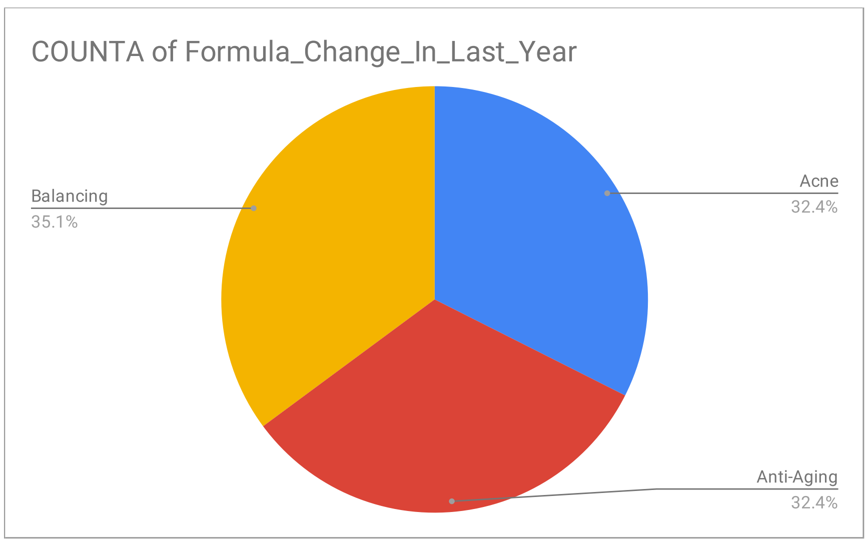
```
select sum(qty) as total_qty_ordered, sum(qty*unit_price) as total_revenue
from Orders
where order_date >= '2015-07-01' and order_date < '2015-08-01'
group by prod_group;
```

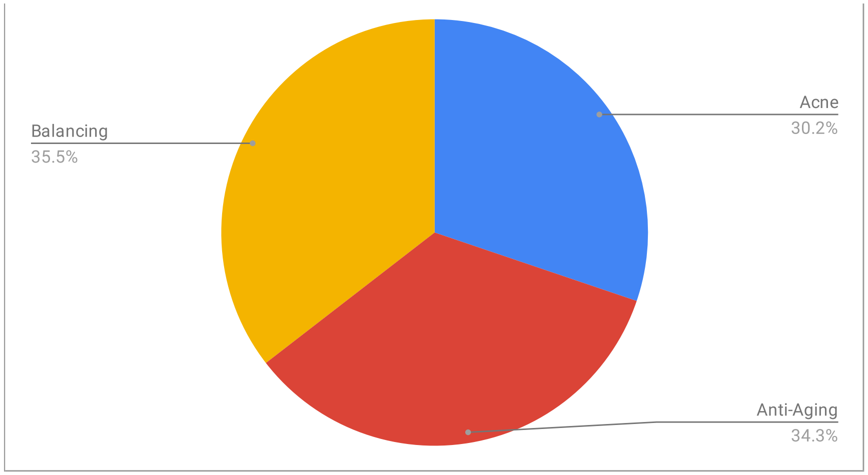
customer_id	retention_days	formula_changes	current_formula_type	Formula_Change_In_Last_Year
57602	88	2	Anti-Aging	Yes
44257	81	2	Balancing	Yes
63100	15	3	Balancing	Yes
71250	123	2	Balancing	Yes
53005	52	2	Acne	Yes
46809	91	0	Acne	No
42722	40	2	Anti-Aging	Yes
89771	68	2	Balancing	Yes
60547	63	3	Acne	Yes
25225	104	0	Acne	No
39925	37	1	Balancing	Yes
43813	146	0	Acne	No
33791	23	2	Balancing	Yes
73401	95	1	Acne	Yes
85803	125	1	Acne	Yes
93040	56	1	Balancing	Yes
33698	126	0	Anti-Aging	No
29584	92	0	Balancing	No
53624	100	0	Acne	No
64086	47	2	Anti-Aging	Yes
15995	117	3	Acne	Yes
28843	104	0	Acne	No
55781	81	1	Anti-Aging	Yes
95341	52	3	Anti-Aging	Yes
38149	7	0	Balancing	No
14319	11	1	Anti-Aging	Yes
54126	55	1	Acne	Yes
92911	117	0	Acne	No
21586	118	3	Anti-Aging	Yes

85026	39	1	Acne	Yes
46688	51	1	Anti-Aging	Yes
95912	25	2	Acne	Yes
77666	97	1	Acne	Yes
30923	12	0	Acne	No
93195	74	3	Balancing	Yes
58828	56	2	Anti-Aging	Yes
82552	15	2	Acne	Yes
97115	113	0	Balancing	No
41897	67	3	Anti-Aging	Yes
52893	42	2	Balancing	Yes
98733	130	1	Balancing	Yes
92714	66	3	Balancing	Yes
71192	121	3	Balancing	Yes
70259	70	3	Acne	Yes
43305	59	0	Acne	No
91974	87	2	Anti-Aging	Yes
64671	143	3	Balancing	Yes
11006	25	2	Acne	Yes
93708	134	3	Anti-Aging	Yes
43586	13	0	Balancing	No

current_formula_type	AVERAGE of retention_days	COUNTA of Formula_Change_In_Last_Year	AVERAGE of formula_changes
Acne	64.83333333	12	1.833333333
Anti-Aging	69.33333333	12	2.083333333
Balancing	75.30769231	13	2.153846154
Grand Total	**69.97297297**	**37**	**2.027027027**



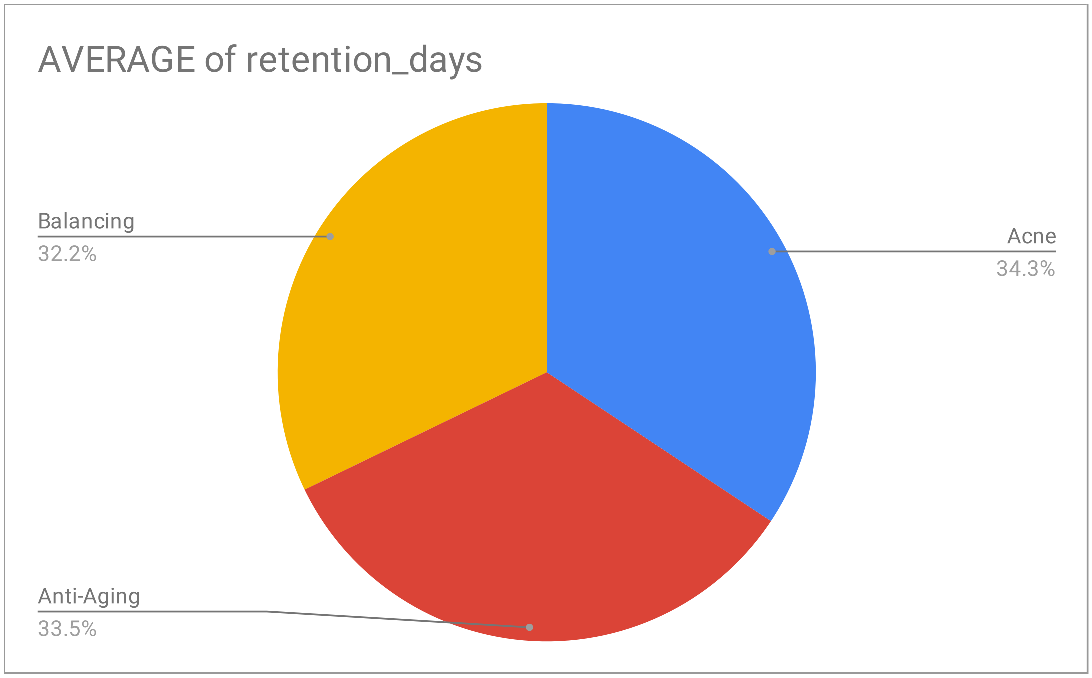




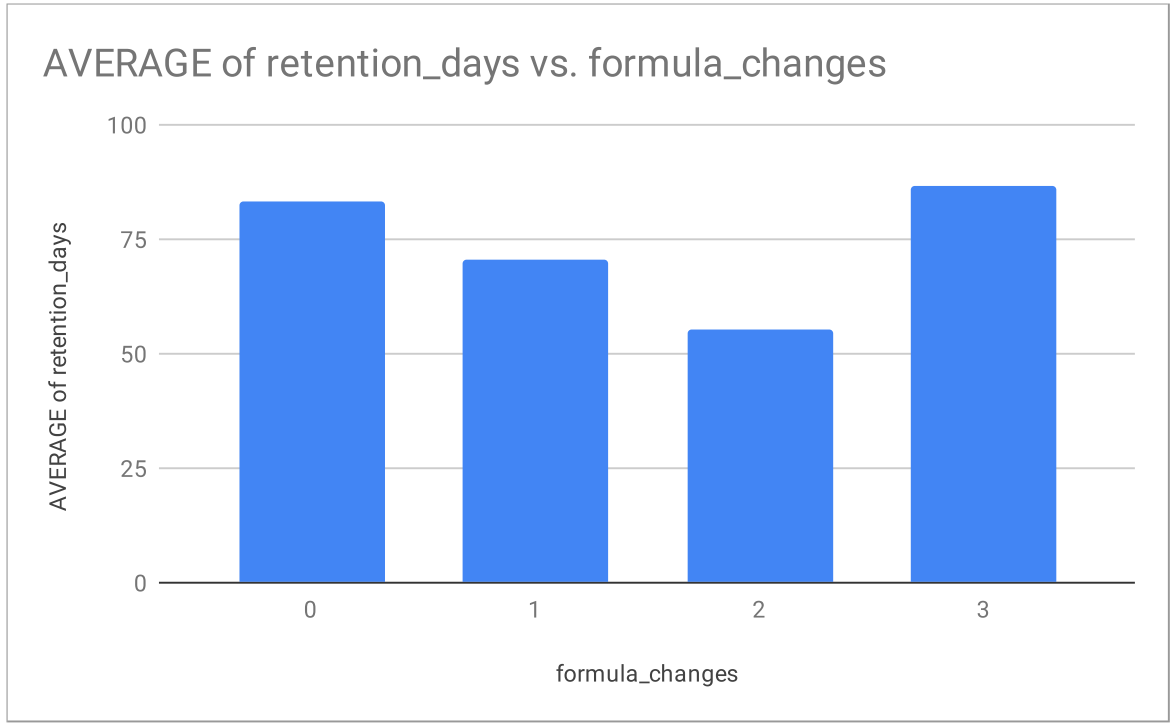
Conclusion I can draw from this analysis is that regardless of formula type, there is not much difference (even sections in pie charts) in: those who changed formula in last year and formula changes. However, there is difference in retention days between formula types. Customers seems to stay longest on Balancing, than Anti-Aging and the least time spent on a formula of all three is Acne formula.

For further analysis though, I would like to see a longer time span of data, with more customers, where age is specified since some of the formula are age-dependent. Also, it would be interesting to see data on seasons within a year when purchases are made since skin care can be depedent upon weather conditions as well. Lastly, data on whether or not there were promotions given in order to incentivize retention on product would be good data points to have to perform further analyses on formula types, etc.

current_formula_type	AVERAGE of retention_days
Acne	75.55
Anti-Aging	73.69230769
Balancing	70.82352941
Grand Total	**73.46**



formula_changes	AVERAGE of retention_days
0	83.38461538
1	70.63636364
2	55.14285714
3	86.66666667
Grand Total	**73.46**



retention_days
88
81
15
123
52
91
40
68
63
104
37
146
23
95
125
56
126
92
100
47
117
104
81
52
7
11
55
117
118

Summary Statistics on retention_days:

Mean	73.46
Median	69
Mode	81
Standard Dev	39.68406353
Variance	1574.824898
Min	7
Max	146
Range	139

Given these descriptive statistics, customers usually use Curology for 2 months and a half.

This is somewhat close to the halfway point between the min and max amount of retention days.

The most frequent amount of time that a customer is on the platform is close to 3 months.

Assuming Normal distribution of data, about 68% of customers (within 1 standard deviation) are between 33 and 113 days of retention or a month to 4 months onboard.

The shortest amount of time a customer stayed with Curology was a week and the longest amount of time was 4.5 months.

In the end, customers are satisfied or needing to try something new within 7 days to a 4.5 month time frame.

This information would be useful to Marketing and Sales team to understand typical customer lifetime value.

39								
51								
25								
97								
12								
74								
56								
15								
113								
67								
42								
130								
66								
121								
70								
59								
87								
143								
25								
134								
13								

Let's say that we have used the dataset to create a model that predicts whether or not an existing user is likely to churn (cancel their Curology subscription). Once we have identified those users, we can take preemptive action and give the user a free gift in their next shipment.

What is the Type I and Type II error in this case, and which scenario is better or worse?

In this scenario,

Type 1 error:	Consequences:
Customer actually churned, but model predicted that the customer did not churn.	Inflated view of onboard customers. Company is at greater loss than it thinks. Can lose profit quickly and not be quick enough to identify churners to bring them back on platform.
Type 2 error:	
Customer did not actually churn, but the model predicted the customer did churn.	Would just be a waste of money on free gifts. Company thinks they are in worse situation than they really are in.

Usually, Type 1 errors are the worse scenarios in general, especially in the health/medicine context. However, Type 1 versus Type 2 is really a case-by-case assessment. So to determine which Type is worse, need to look at the consequences from each error type.

In this situation, not identifying churn in time and therefore not making soon enough effort to keep customers would be a greater cost to Curology.

This is versus funds mistakenly spent to send free gifts to customers who are actually not churning, but stlll on the platform.