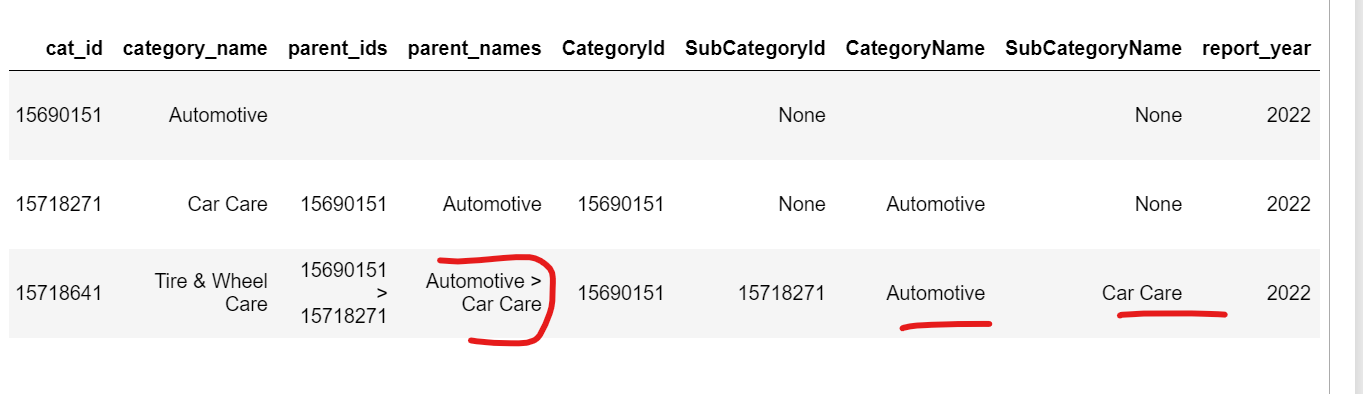
**Q.No.1 Products with the highest market potential in one of the categories and subcategories.**

Ans: For this I splitted parent\_names as “Automotive > Car Care”

Automotive as Category

Car Care as Sub Category.



I found the total sales using: "TotalSales\_360"=dniche\_summary\_maximum\_units\_sold\_t360 \* niche\_summary\_avg\_price. Since

**Average=Sum/No of products.** So I got the sum of sales of each product.

Looking into the below diagram I calculated each top product from every category and subcategory and I found **Home covid Test kit rapid which lies in**

**Category-> Diagonistics and Screening**

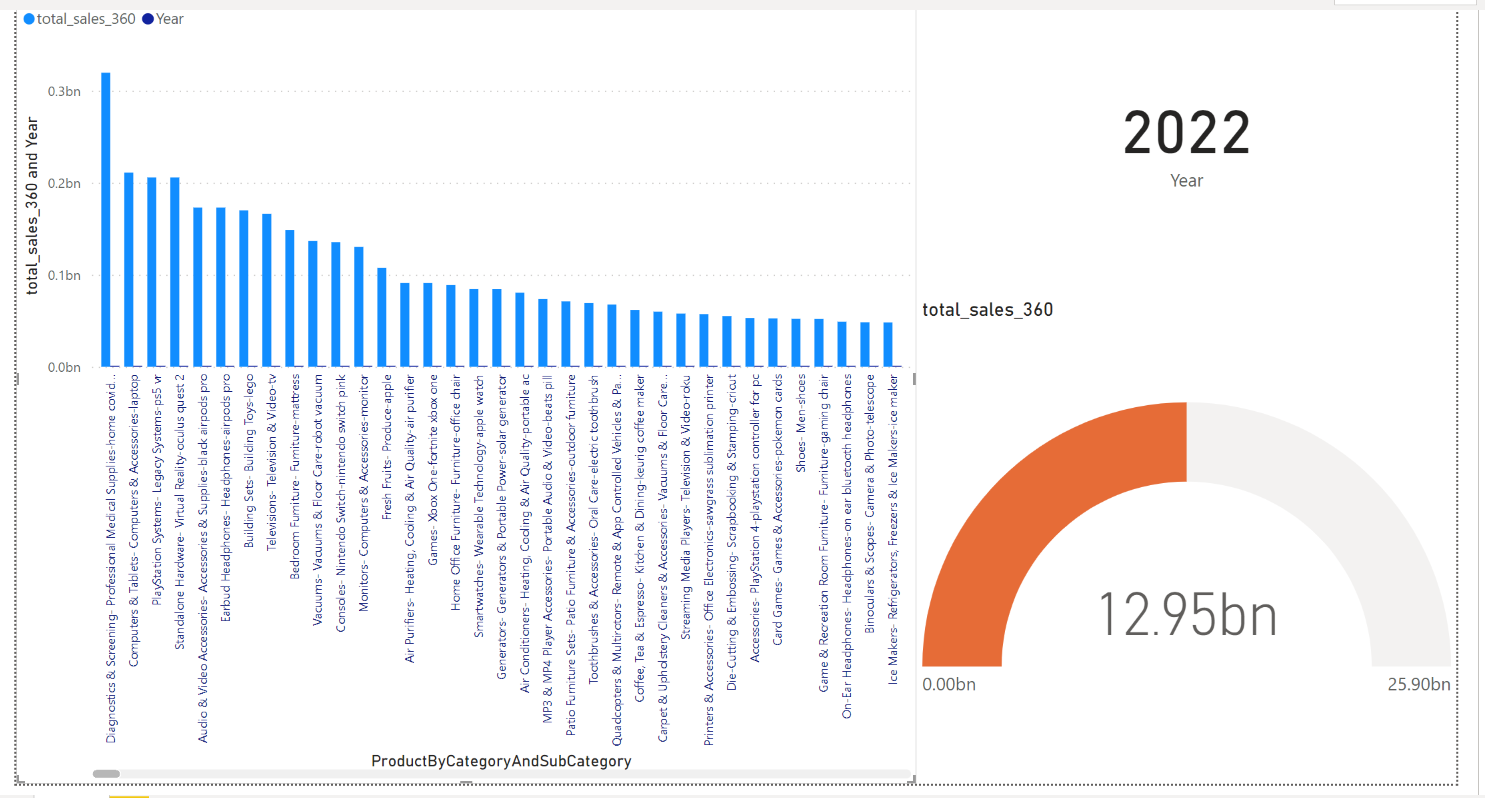
**Sub Category-> Professional Medical Supplies**

maximum sold in **2022-04** which was **319.67** millions dollar sold.

In smartwatch also Apple watch was maximum sold which is **84.44millions dollar**.

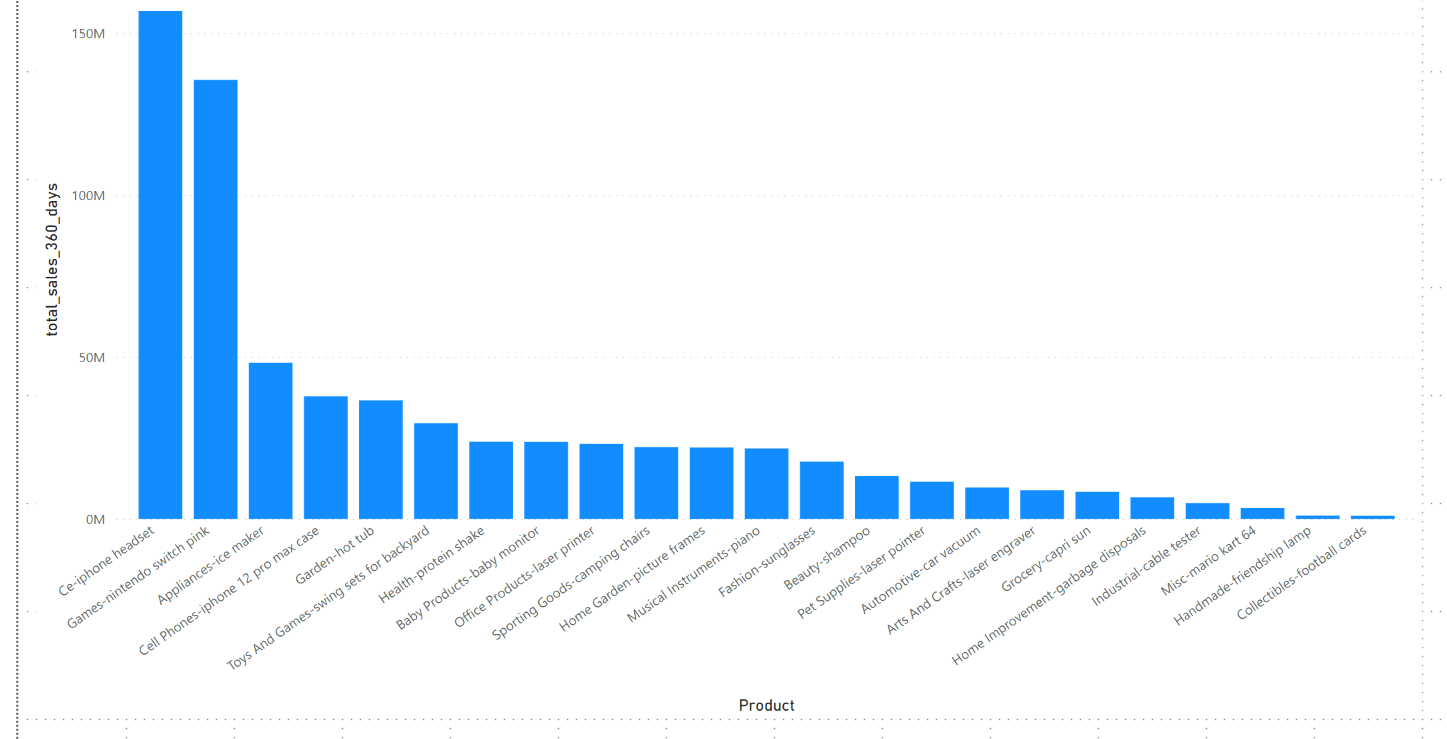
So we can see every top potential product for every category and subcategory.

Please refer the below Diagram:



**Q.No.2 The product that has highest potential across categories.**

Ans: So, for this I considered product which is directly linked to categories:



I found that iphone headset,Nintendo switch pink and ice maker are the product that have most potential among each categories.

**Qno.3 Please find the estimated sales volume for the next year based on the Historical dataset for particular products.**

Ans- I saw that there was no historical Yearly sales data which could have been much easy to analyze.So, I tried using the Mlib of apache spark and used three methods to find the best case:

1.Linear Regression

2.Decision tree regression

3.Gradient-boosted tree regression

I found my r square is 99% while using **linear regression** and but I see the rms error is quite high(due to insufficient training data). Other algorithms have still much high error so I choose linear regression. I considered 70% as training data and 30% as test data.

I consider these **metrics** ['click\_share\_t360', 'click\_count\_t360', 'niche\_summary\_search\_volume\_t360', 'niche\_summary\_maximum\_units\_sold\_t360', 'product\_count\_yoy', 'TotalSales\_360'] as the dependent variables for total sales\_360. Because the product clicks,counts,search volume,maximum units sold and yoy are the parameters that impact total sales.

I found the correlation as:

Correlation to Sales for click\_share\_t360 0.04801700444464939

Correlation to Sales for click\_count\_t360 0.6106991824499717

Correlation to Sales for niche\_summary\_search\_volume\_t360 0.697499141592546

Correlation to Sales for niche\_summary\_maximum\_units\_sold\_t360 0.4999795785990298

Correlation to Sales for product\_count\_yoy 0.02857406905565327

Correlation to Sales for TotalSales\_360 1.0

My report looks like this after the analysis:

Pink represents **sales\_2022**.

Purple represents **Sales\_2023**.

