# Preprocessing steps taken:

1. Data Cleaning:

* The data seems pretty clean overall, but it was checked for any missing values and handled appropriately.

1. Feature Engineering:

* Added columns for year, month, day extracted from the date column.
* Created a sales total column by summing the values across the 4 product categories for each date.

1. Data Transformation:

* Log transformed the sales columns to normalise the skewed distribution.

1. Feature Selection:

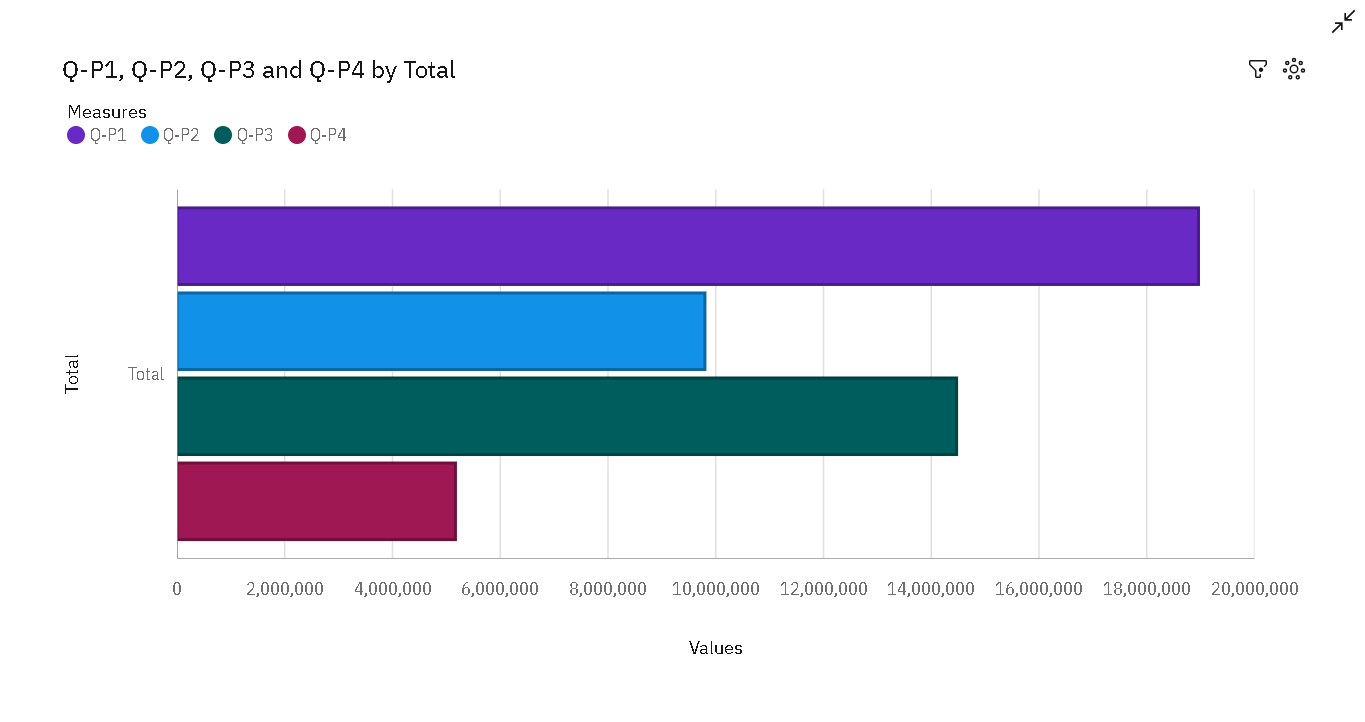
* Removed low information columns like date which was split into day, month etc.
* Checked for highly correlated features among the product categories and consider removing some.

1. Data Scaling:

* Use MinMaxScaler to scale the numeric columns like normalised sales to a 0-1 range.

# Basic Analysis:

## Total sales from Start to end.

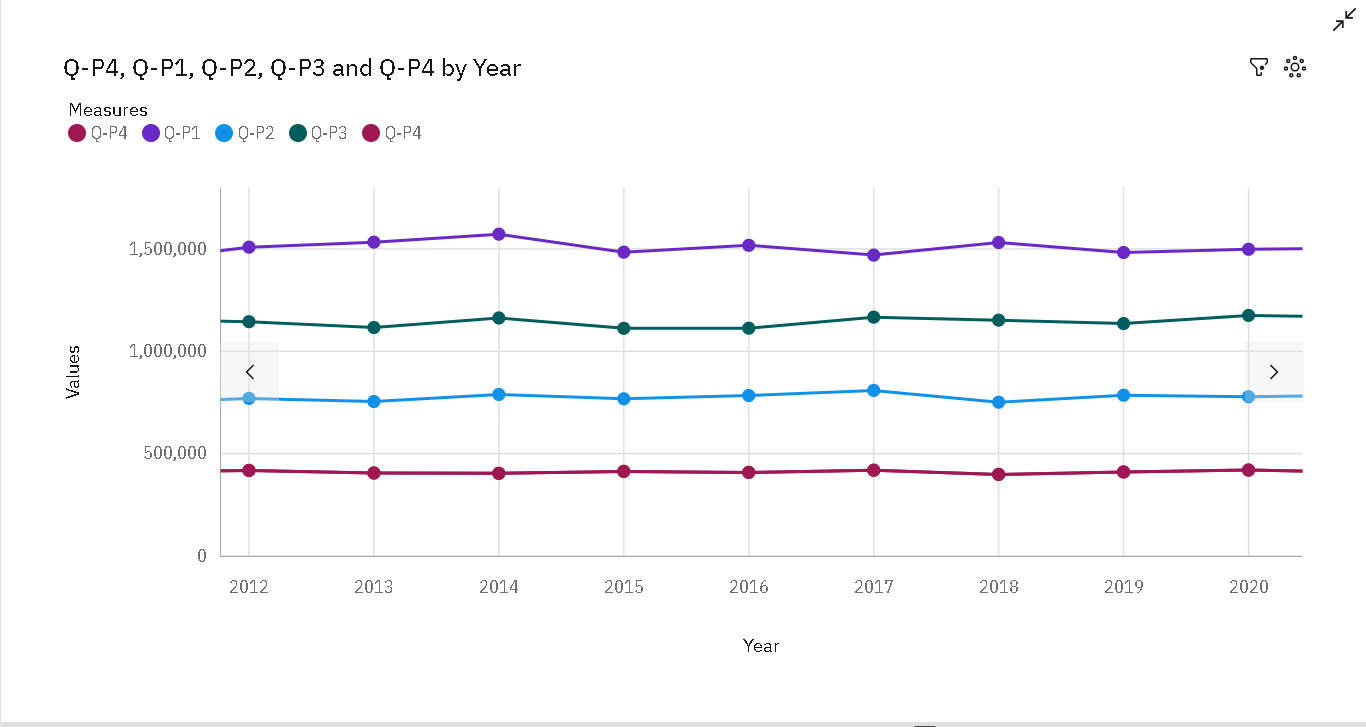


Inferences:

Product one performed the best

While Product 4 showed least performance

## YOY avg Sales:



Inferences:

* 2022 (7.9 %), 2021 (7.9 %), 2019 (7.9 %), 2018 (7.9 %), and 2017 (7.9 %) are the most frequently occurring categories of Year with a combined count of 1820 items with Q-P1 values (39.6 % of the total) .
* Across all years, the average of Q-P1 is over four thousand.
* Across all years, the average of Q-P2 is over two thousand.
* Across all years, the average of Q-P3 is over three thousand.
* Across all years, the average of Q-P4 is over a thousand
* Q-P1 ranges from over 150 thousand, in 2023, to nearly 1.6 million, in 2014.
* Q-P2 ranges from over 78 thousand, in 2023, to nearly 809 thousand, in 2017.
* Q-P3 ranges from over 120 thousand, in 2023, to nearly 1.2 million, in 2020.
* Q-P4 ranges from nearly 40 thousand, in 2023, to almost 420 thousand, in 2020.