Module 2 - Data Analysis of sales USA

Questions 5 to 10

The Data product_a.csv consists of 14 columns including index column and total sales per week with weekly Date feature. It has got 18,249 rows with no null values.

Date feature (weekly basis) is crucial here as to check the time series trend of sales volume or price. In this dataset, the column Total_vol is the total of plu1, plu2, plu3 and bags_t.

As per the summary statistics, total sales have multiple modes (Mode is the value of most frequent item) and the feature price is lying between 0.44 and 3.25.

Skewness is a measure of symmetry and a symmetrical dataset will have a skewness equal to 0. skewness = 3*(mean-median)/standard deviation.

Here, the price feature is closer to symmetric and others are highly positively skewed. This is a clear indication of outliers and removal of this issue would provide better results for predicting values.

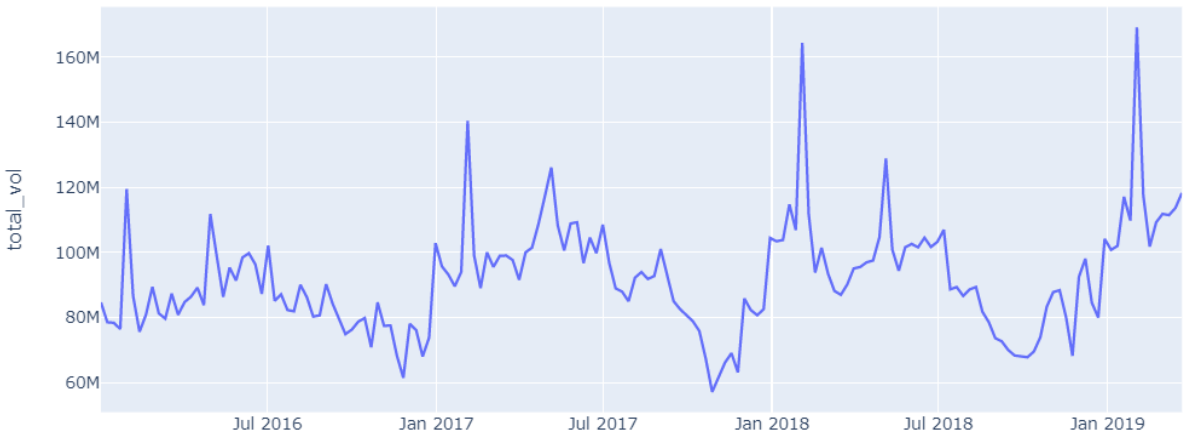
The values of 50th percentile (2nd quartile) and the median are the same. Standard deviation is a good measure of how spread out the numbers are from mean value. Here, bags_t and bags_s has got a high deviation and other variables are closer to their mean values.

As per 'Pearson' correlation (strength and direction of the linear relationship) matrix of this dataset, the features plu1, plu2, plu3, bags_t and bags_s are positively highly correlated to each other. The total_value is highly correlated to plu1, plu2, plu3 and bags_t since the total_vol is the cast of those variables. There is no correlation between price and other variables.

Questions 10 to 12

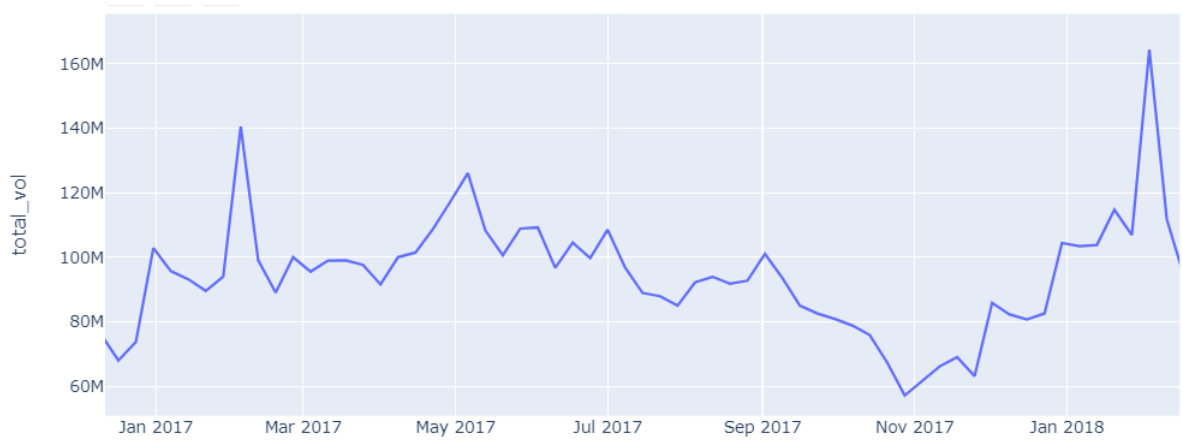
This Data was included with TotalUSA rows as this would be a duplicate when we analyse Total_volume per week. So, these rows have been removed.

Sales Volume (total) per year



As per the above time series plot, the trend of sales volume increases from year 2016 to 2019 despite the fluctuation throughout the months. The sales dropped in the end of year 2017 and it reached the peak in the beginning of year 2019.

Sales volume (total) per month



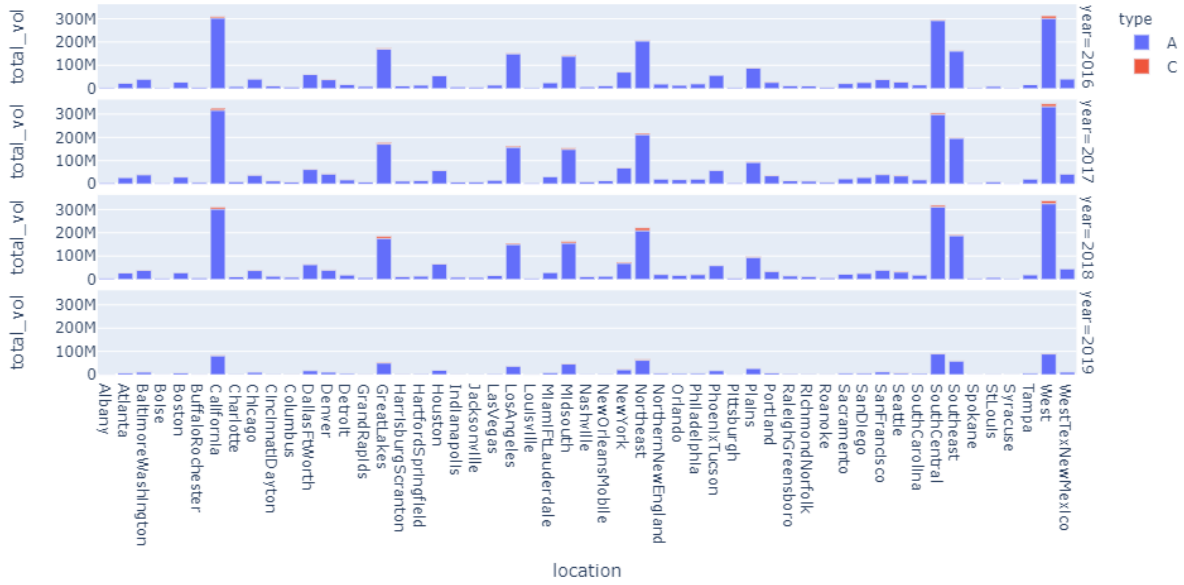
As per the time series plot of sales volume per month, it shows the sales volume goes to the highest point in each year in the month of February and fluctuate towards the end of the year. It goes to the lowest point in November each year.

Price vs time

The above plot shows a clear indication that the price goes to the lowest point in February and it reaches the peak in November.

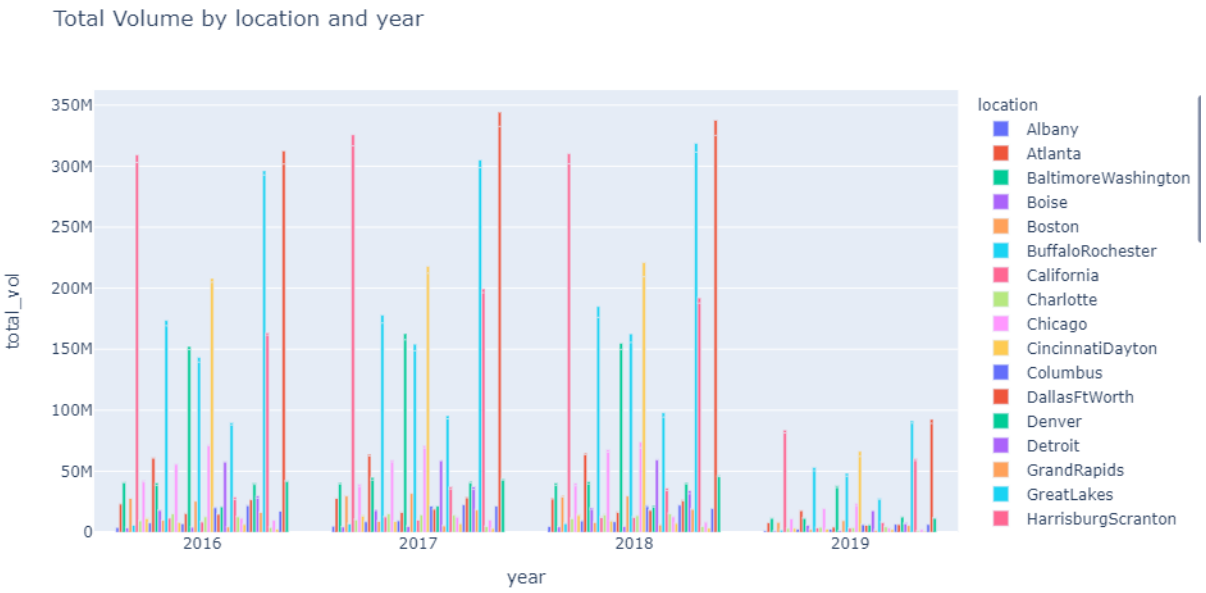
Overall, there is a negative effect (relationship) between the Total volume and Price.

The Bar plot of Total_volume per location by type



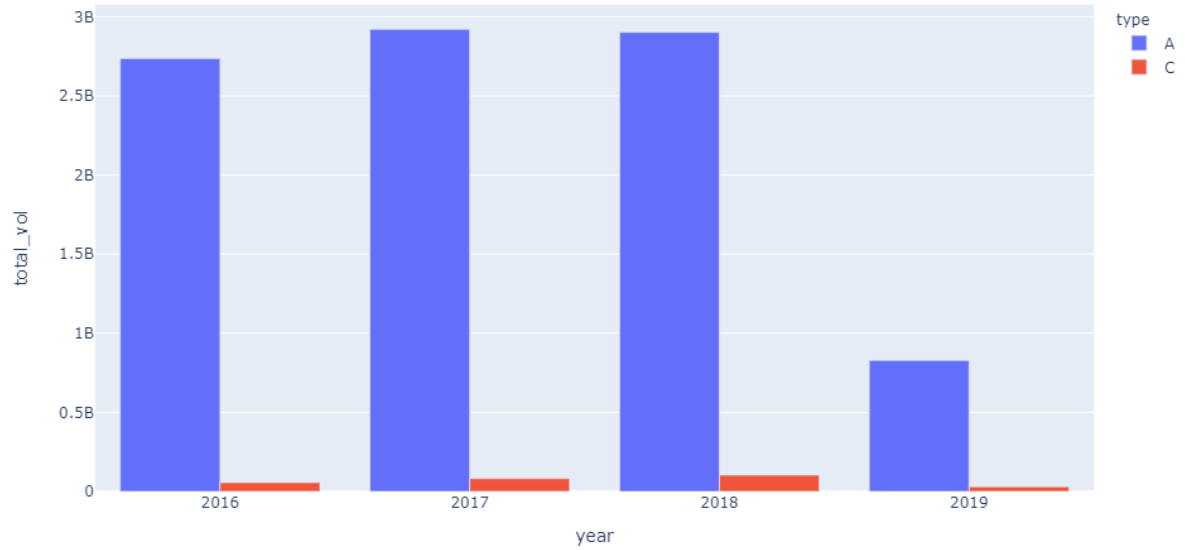
A per the above plot, the sales of type A is nearly 300M which is extremely high compare to type C and California, West and South central have got the highest volume of sales throughout the years 2016 to 2018 but it dropped in 2019 compared to the previous years.

The Bar plot of Total_volume per location



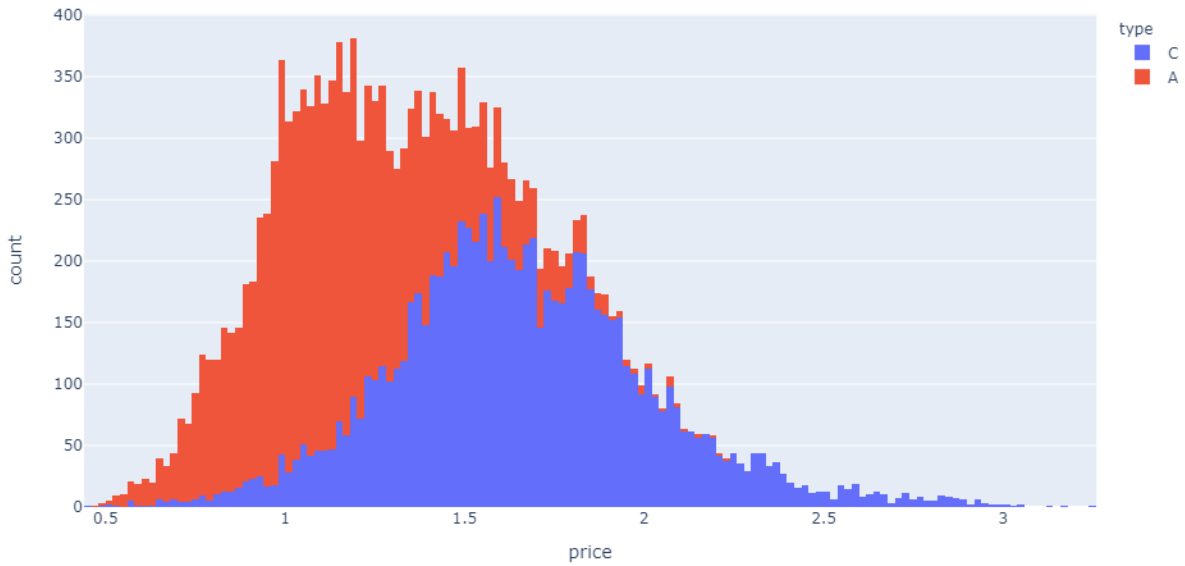
The above plot clearly shows the locations that have got highest sale volume are the same as above mentioned locations in the previous plot. The sales volume dropped in 2019 dramatically while having the same highest sale volume happened in the same location as the previous years.

The Bar plot of Total_volume per type



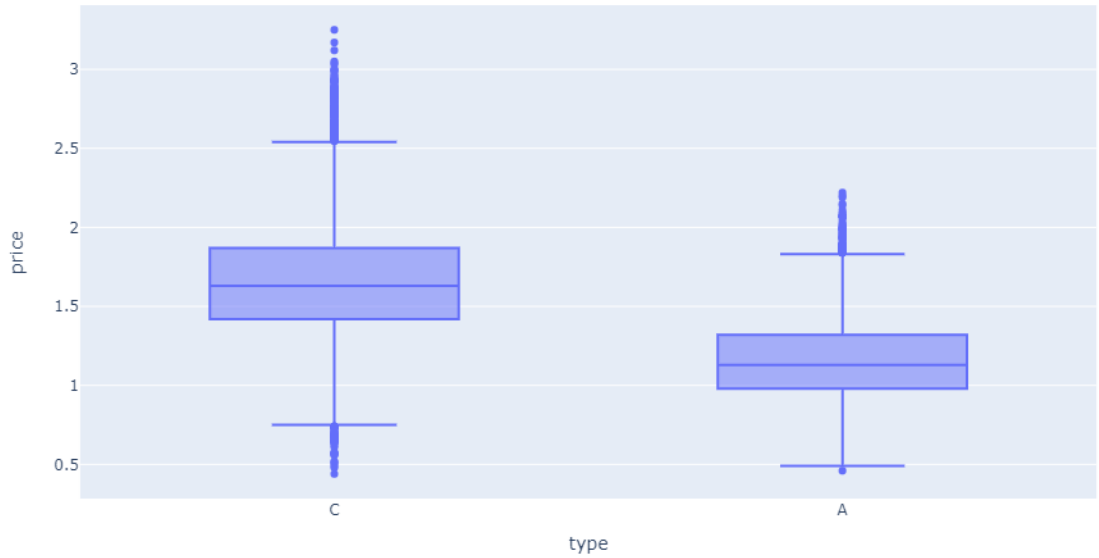
This bar plot explains the sales volume dropped in 2019 and the proportion of A is to B is extremely high. Type C sales volume per year never exceeded 0.2 Billion whereas Type A total volume per year used to be nearly 3 Billion except for 2019.

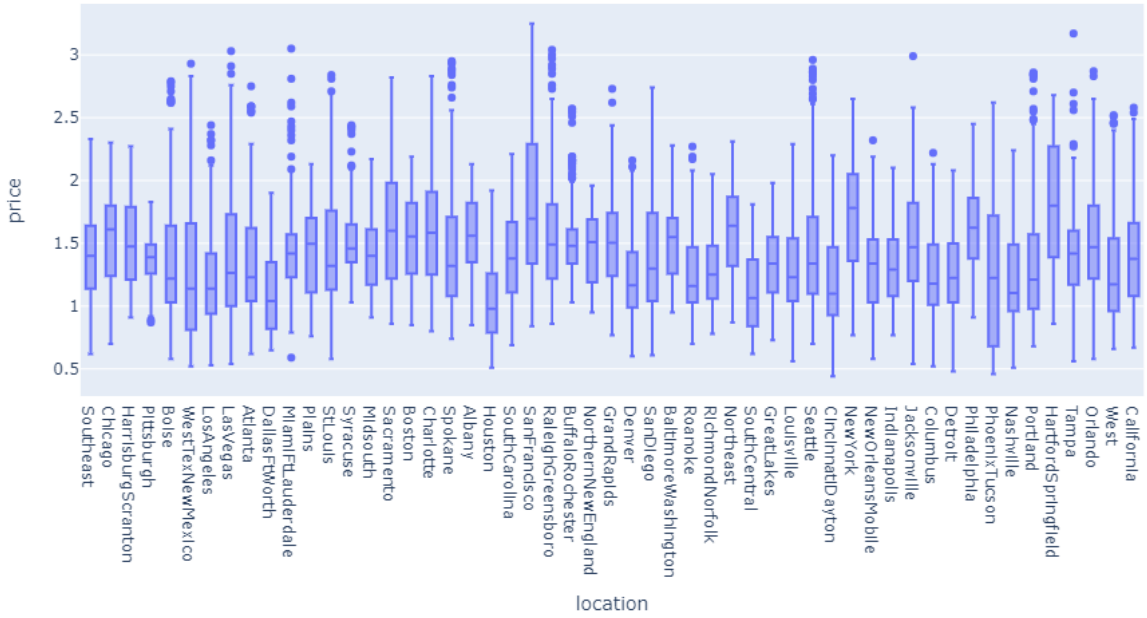
Histogram of Price



The above histogram clearly shows the price distribution of both A and B are closer to symmetric and lying in a bell curve. Type A prices are lower then type C but the bell curve is flat for A compare to type C.

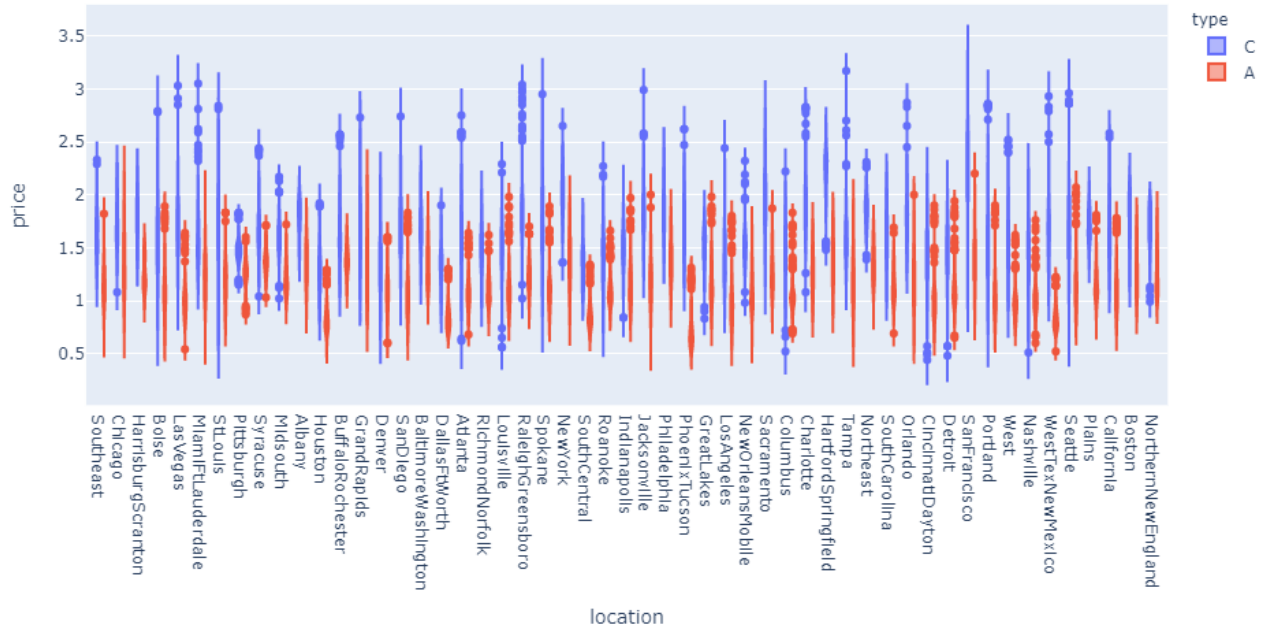
Box plots of Price





The prices of items A and B are varying location to location and many of them have got outliers. As the price distribution varies for locations the mean value of the same varies as well. It is impossible to determine a constant value for price per location.

Box plot of Price per type



It is another clear visual of price discrepancy of type A and Type C and type A is lower than C.

St louis, San Francisco and Detroit have got a widely spread distribution of price for type C.