Group Name

J & H Group

Members

1) Hina Merab Asif
    - Email: hinamerabtimothy@gmail.com
    - Country: England
    - Specialization: Data Science
2) Jasmine Luo:
    - Email: jazlallinone@gmail.com
    - College: University of Southern California
    - Specialization: Data Science
3) Junfei Liu:
    - Email: junfeiliu.jeff@gmail.com
    - College: University of Rochester
    - Specialization: Data Science

Problem Description

A large beverage company in Australia sells its products for a whole year, and the selling of those products is influenced by some factors like seasons or holidays. They need a weekly forecast for each of their products, which can predict how much of a product will sell in the current week.

Data Understanding

a) What type of data have you got for analysis?

   Both numerical and categorical data. We have three numerical attributes: sales, price discount, and Google mobility, among which sales is the outcome we want to predict. The rest nine attributes are categorical, including product consisting of SKU types from 1 to 6, date of the start of a week from 2/5/2017 to 12/27/2020, and In-Store Promo, Catalogue Promo, Store End Promo, Covid_Flag, V_DAY, EASTER, CHRISTMAS as boolean values.

b) What are the problems in the data (number of NA values, outliers , skewed etc)?

   Field Summary Table

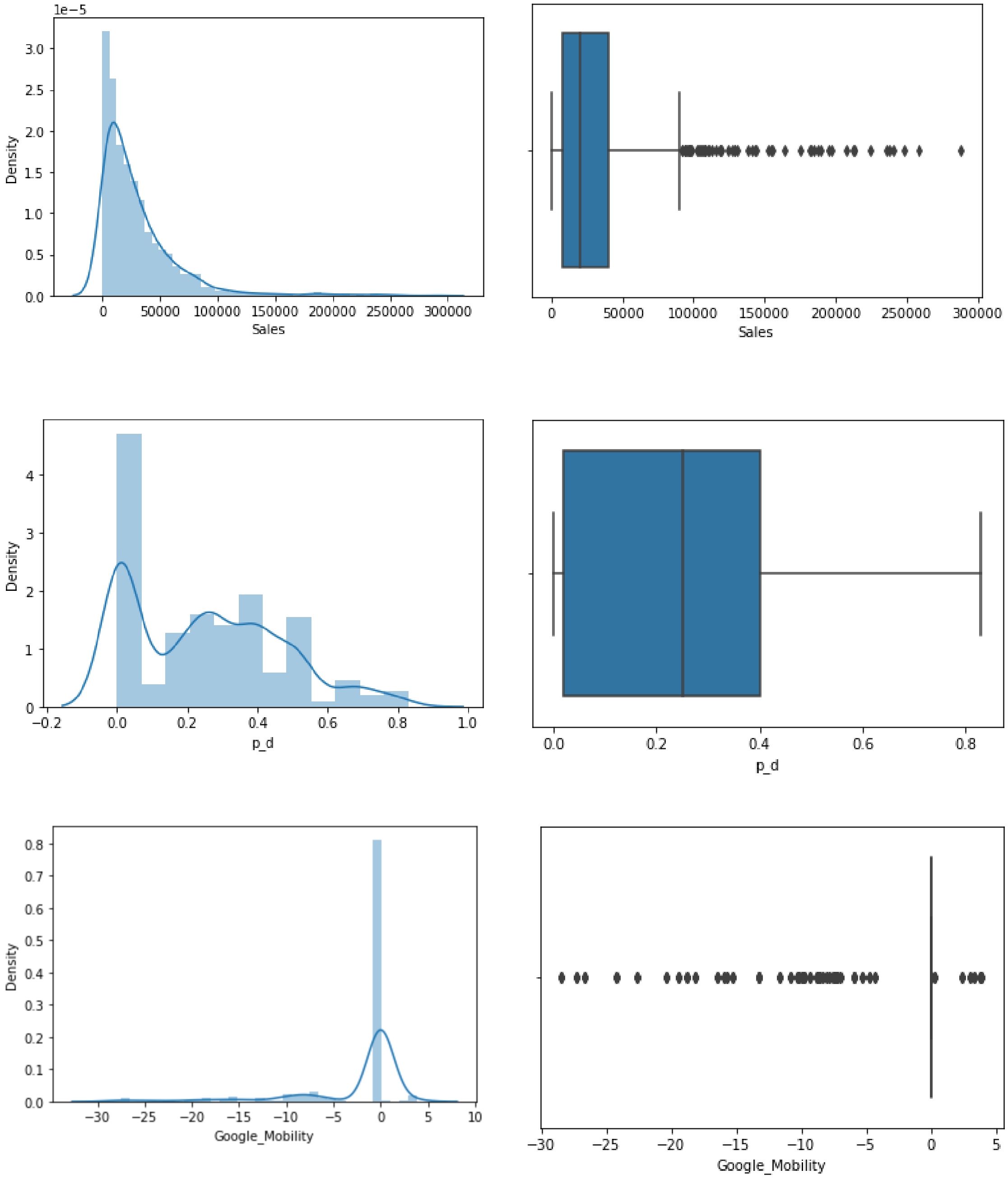
1) Numeric field

Name	**%Populated**	**Min**	**Max**	**Mean**	**Standard Deviation**	**%Zeros**
Sales	100%	0	288322	30294.68	35032.53	2.46%
Price Discount	100%	0	0.83	0.25	0.22	18.97%
Google_ Mobility	100%	-28.49	3.9	-2.38	5.81	77.34%

2) Categorical field

Name	**%Populated**	**#Unique Values**	**Most Common Field Value**
Product	100%	6	SKU1, SKU2 , SKU3, SKU4, SKU5, SKU6
date	100%	204	NA
In-Store Promo	100%	2	0
Catalogue Promo	100%	2	0
Store End Promo	100%	2	0
Covid_Flag	100%	2	0
V_DAY	100%	2	0
EASTER	100%	2	0
CHRISTMAS	100%	2	0

For all numerical fields, there is a skewness problem. Besides the Price Discount field, Sales and Google_Mobility fields have the problem of outliers. Also, there is a percentage of zeros in the Sales field, which may need further manipulation. For the other two fields, a relatively large proportion of zeros will not affect further works.



c) What approaches are you trying to apply on your data set to overcome problems like NA value, outlier etc and why?

There are no NA values in our dataset. Outliers will temporarily not be removed because sales and mobilities could be fairly higher than average in reality.

Zeros in Google mobility are not concerning because the mobility data is only valid when the Covid flag is true.

For zero values in Sales, we noticed that only the sales data for SKU1-5 from 11/22/2020 to 12/27/2020 are zero and the data for SKU6 in this time period is missing. It is reasonable to believe that the sales data during this time period is abnormal and should not be generalized. However, it is arguable to say that zero sales in this period are because the date is closely related to sales data. Therefore, whether data entries within this period should be removed remains controversial.

Apparently, training models with these data entries included will possibly result in overfitting and cause unsatisfying accuracy on test data. Thus, we decide to remove them now but leave the option to include them again.

A complementary plan would be to train a separate model for this specific period, yet the predictability is not promising due to few training data.