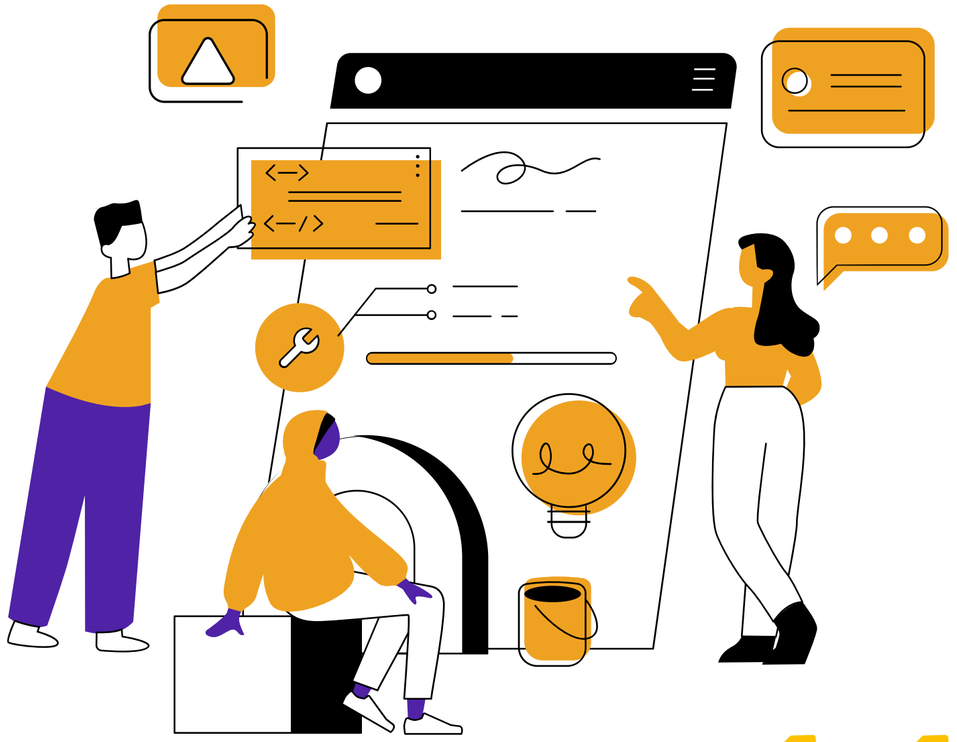
Forecasting

Sales of product at Favorita stores located in Ecuador

By Ni Nyoman Triwahyuni

Source : https://www.kaggle.com/competitions/store-sales-time-series-forecasting/data



ABOUT

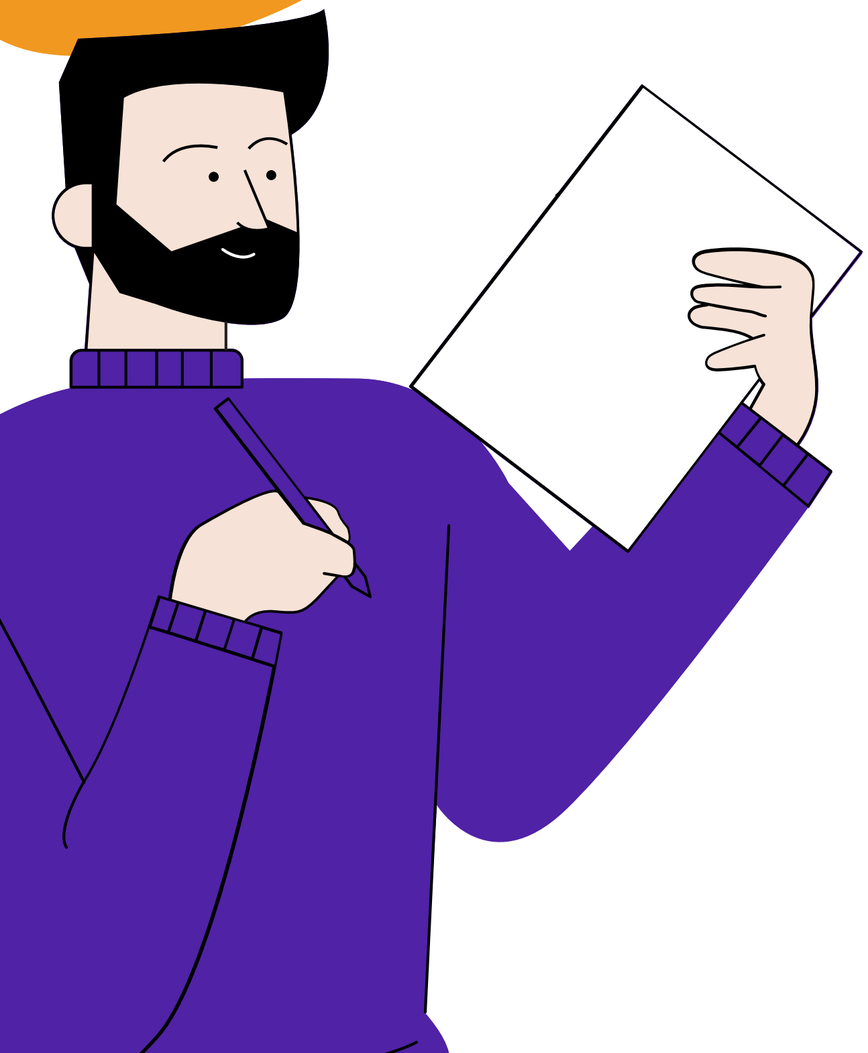
INTRODUCTION

Because of the fluctuative oil price nowadays, therefore need to explore how oil price affect sales of the product. In this case, we use sales data at Favorita stores located in Ecuador which very sensitive of the oil price.
Here are the dataset :

- train.csv
- sample_submission.csv
- stores.csv
- oil.csv
- holidays_event.csv

AGENDA

- Objectives
- Data Preparation
- Result
- Business Solution
- References



Objectives

1.Standard Data Explarotary

- Statistical Summary
- Univariate Analysis
- Multivariate Analysis

2. Deep-dive Data Exploratory

- How the growth of sales, product available on promotion and oil price?
- How the correlation between sales, onpromotion and oilprice in monthly basis?
- How the day type affect sales ?
- Who is the top 10 customer who bought product the largest ?
- How the growth of the sales in each city ?

3. Model of Forecasting Sales

4. Business Solutions

Data Preparation

1.Import libraries & Dataset

2. Data Cleansing

- Handling Missing Value
- Handling Duplicate Value

3. Data Manipulation

- Filter data in range date from Jan 1st, 2013 till Jul, 31th, 2017
- Merge dataframe
- Rename "day_type"
- Fillna mean of dcoilwtico
- Handling Outliers - Clipping Method (for Modelling only)
- Label One Hot Encoding (for Modelling only)
- Split Data (Train - Validate - Test Data for Modelling only)

Dataset ready !

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3054348 entries, 0 to 3054347
Data columns (total 13 columns):
 #   Column       Dtype
---  ------       -----
 0   id           int64
 1   date         object
 2   store_nbr    int64
 3   family       object
 4   sales        float64
 5   onpromotion  int64
 6   city         object
 7   state        object
 8   type_x       object
 9   cluster      int64
 10  day_type     object
 11  description  object
 12  dcoilwtico   float64
dtypes: float64(2), int64(4), object(7)
memory usage: 326.2+ MB
```

RESULT

1. Standard Data Explarotary

Standard Data Exploratory

Statistical Summary

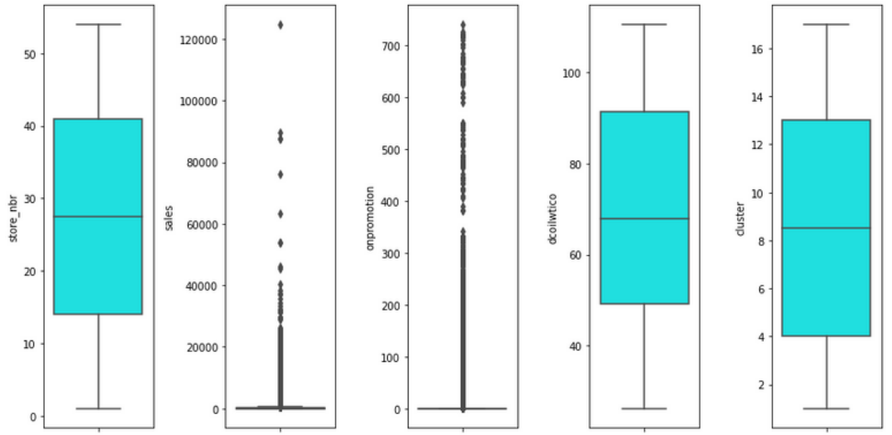
Numerical Variables

	store_nbr	sales	onpromotion	dcoilwtico	cluster
count	3.054348e+06	3.054348e+06	3.054348e+06	3.054348e+06	3.054348e+06
mean	2.750000e+01	3.590209e+02	2.617480e+00	6.801587e+01	8.481481e+00
std	1.558579e+01	1.107286e+03	1.225494e+01	2.129874e+01	4.649735e+00
min	1.000000e+00	0.000000e+00	0.000000e+00	2.619000e+01	1.000000e+00
25%	1.400000e+01	0.000000e+00	0.000000e+00	4.910000e+01	4.000000e+00
50%	2.750000e+01	1.100000e+01	0.000000e+00	6.801587e+01	8.500000e+00
75%	4.100000e+01	1.960110e+02	0.000000e+00	9.153000e+01	1.300000e+01
max	5.400000e+01	1.247170e+05	7.410000e+02	1.106200e+02	1.700000e+01

Categorical Variables

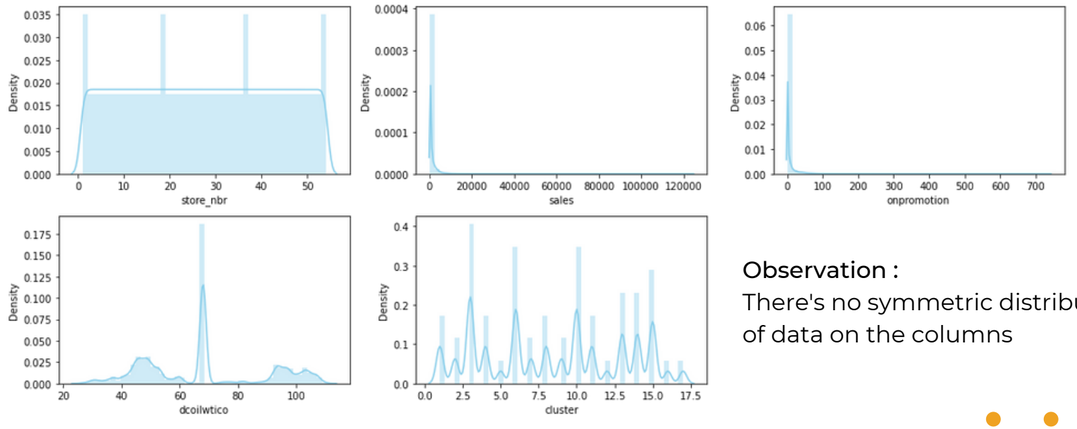
	family	city	state	day_type
count	3054348	3054348	3054348	3054348
unique	33	22	16	7
top	AUTOMOTIVE	Quito	Pichincha	Normal
freq	92556	1018116	1074678	2551824

Univariate Analysis



Obervation :

1. There is no outlier in the in the column store_nbr, dcoilwtico, and cluster, except sales and onpromotion
2. The data distribution of sales and onpromotion are not symmetric and there's a lot of outlier data.

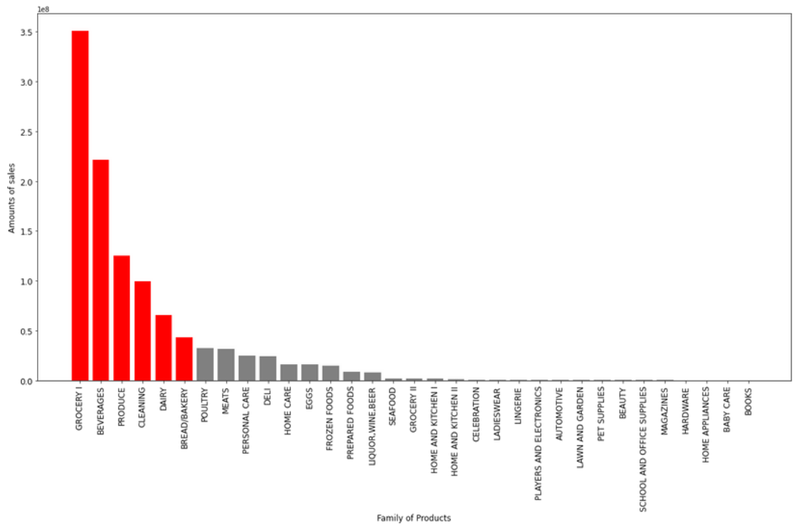


Observation :
There's no symmetric distribution of data on the columns

Standard Data Exploratory

Multivariate Analysis (Categorical Variables)

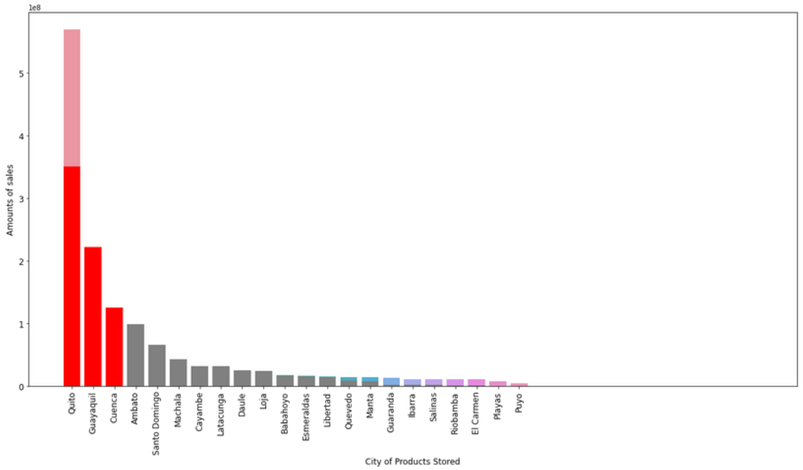
Amount of Sales Family of Product



Family of product which have sales more than average are

1. Grocery
2. Beverages
3. Produce
4. Cleaning
5. Dairy
6. bread/bakery.

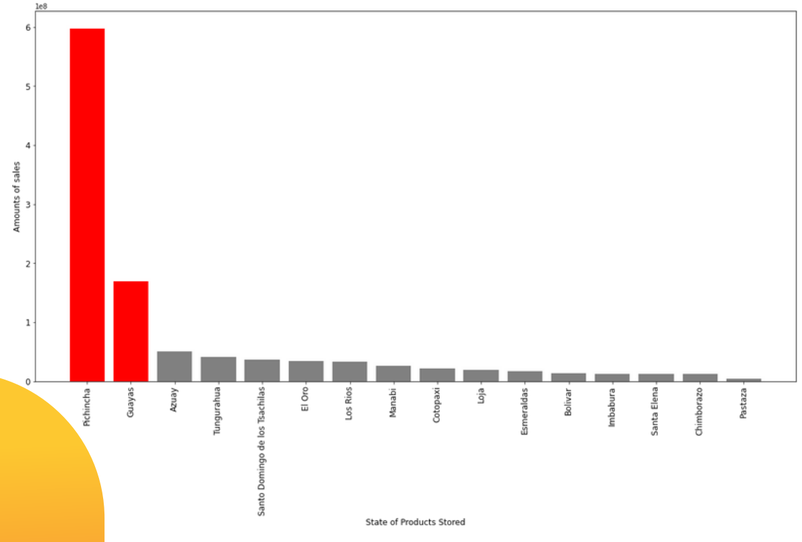
Amount of Sales in City



City which have sales more than average are :

1. Quito
2. Guayaquil, and
3. Cuenca.

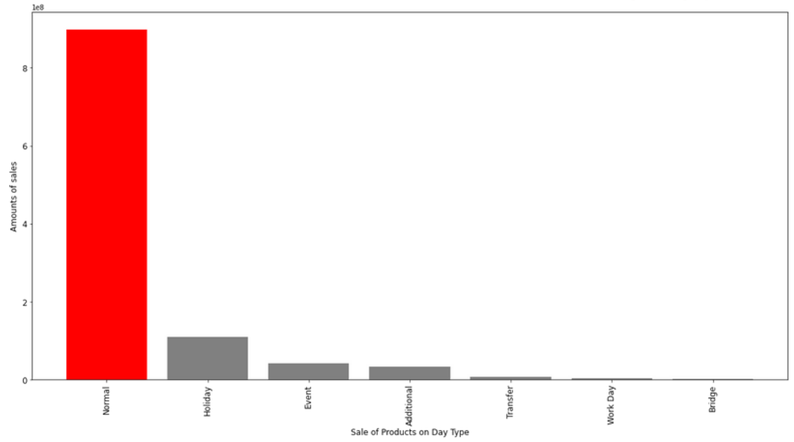
Amount of Sales in State



State which have sales more than average are :

- Pichincha
- Guayas

Amount of Sales in Each Day Type

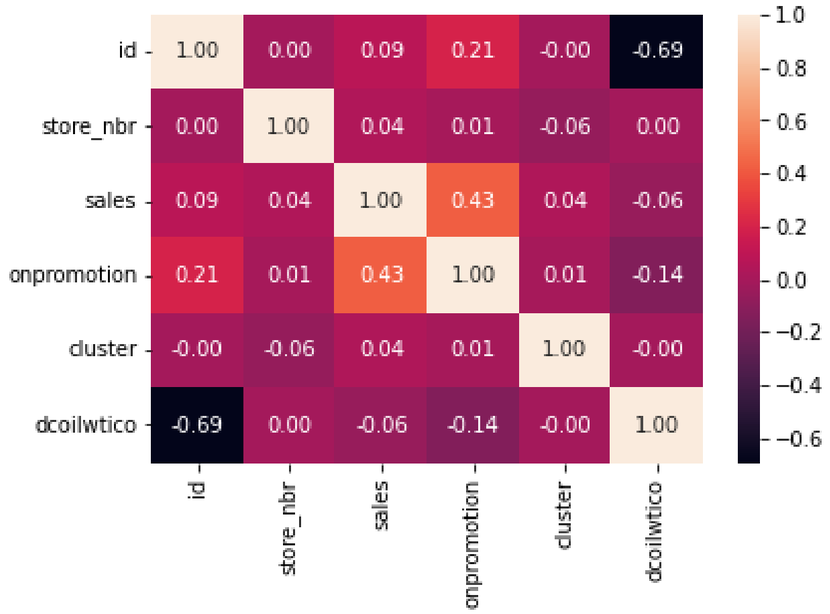


Day type which have sales more than average is on the "Normal" day.

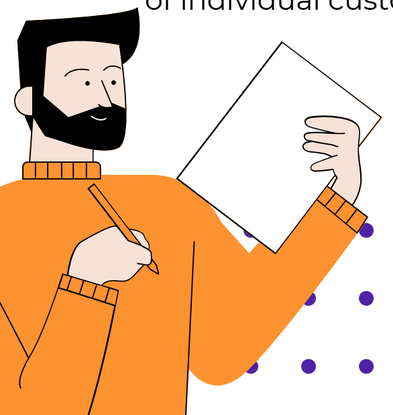
Standard Data Exploratory

Multivariate Analysis (Numerical Variables)

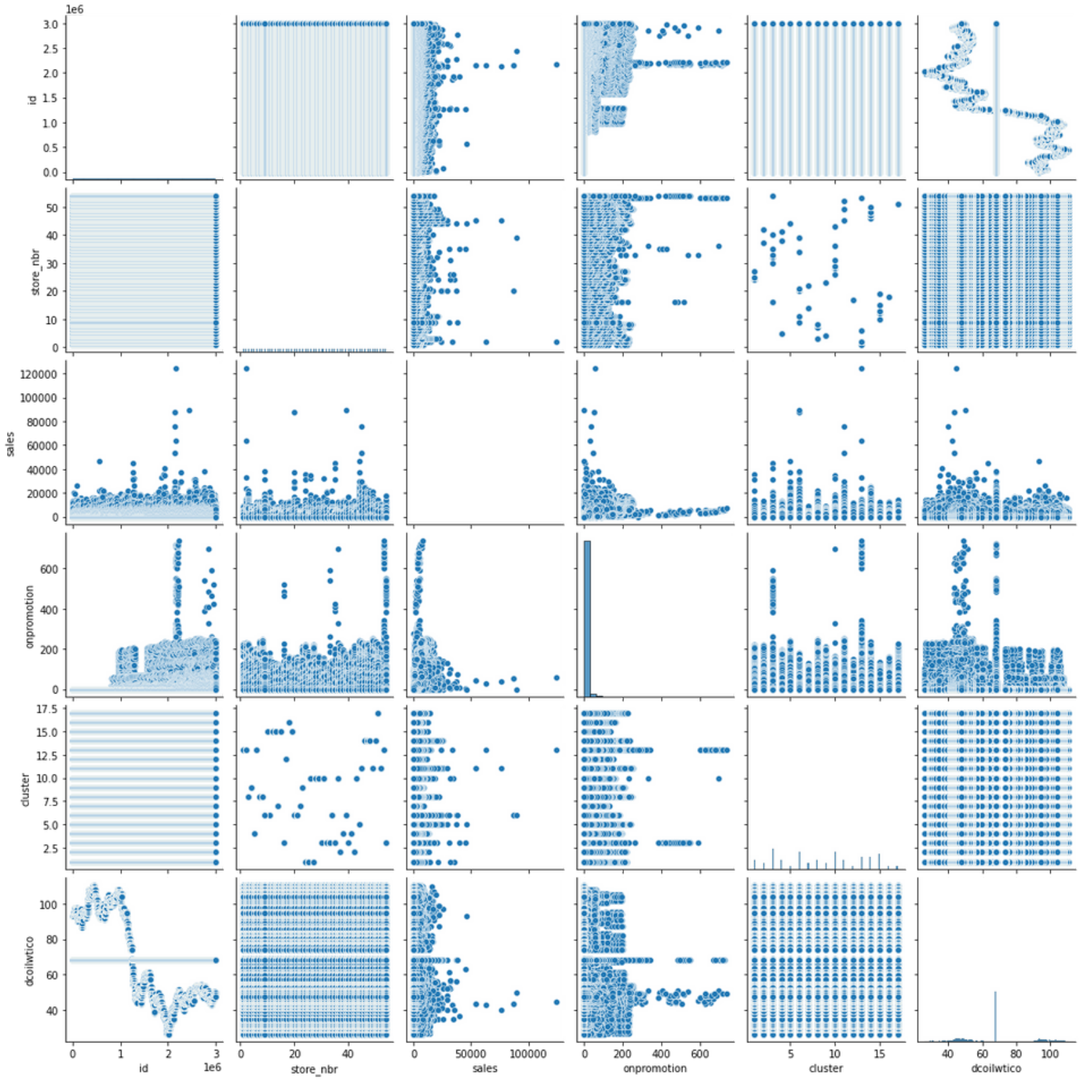
Heatmap Correlation



Observation :
There's no significant correlated between variables of individual customers.



Paiplot of Numerical Variables



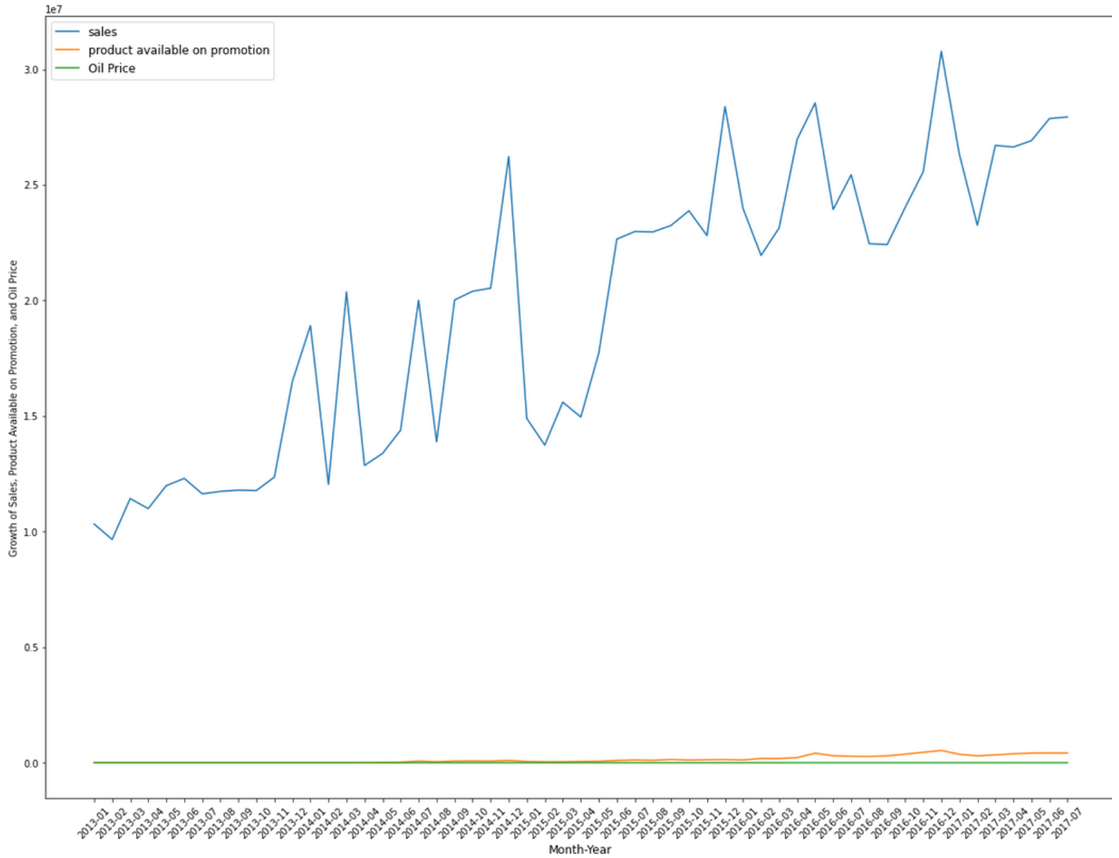
Observation :
From the graph we know that :sales and onpromotion evenly distribute data on store_nbr, cluster and dcoilwtico(oil price).

RESULT

2. Deep-Dive Data Explarotary

Deep-Dive Data Exploratory

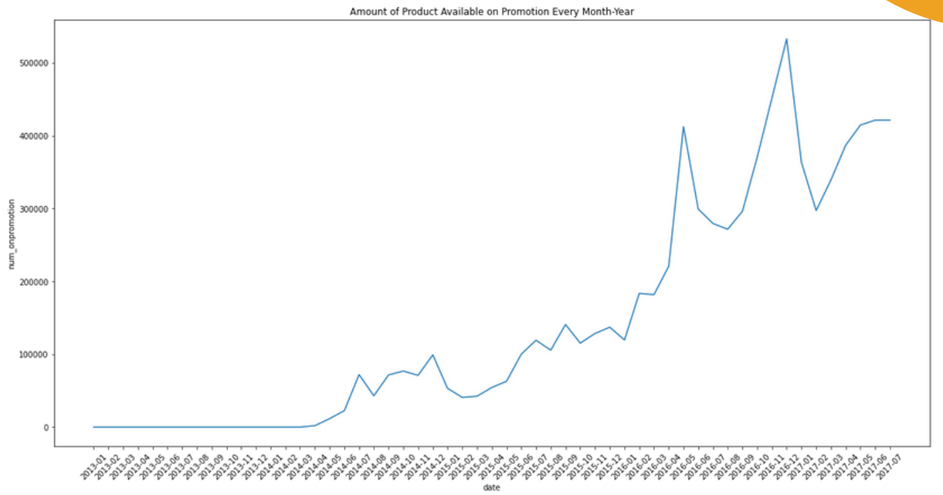
1. How the growth of sales, product available on promotion and oil price?



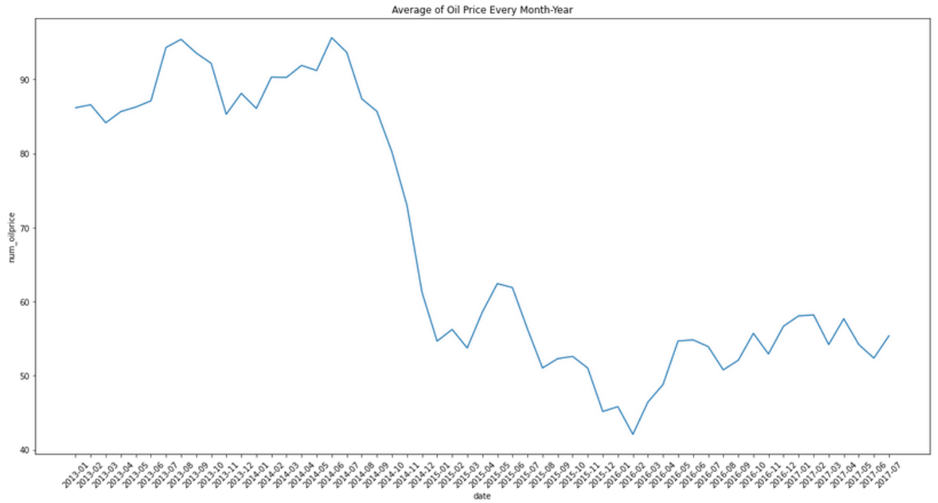
Observation :

amount of sales has significant different of amount than the amount of product available on promotion and oil price, but overall there's still increase of growth sales from Jan 2013 till Jul 2017. It's inline with the increase product available on promotion and decrease of oil price which indicate Ecuador really sensitive with oil price.

How the growth of product available on promotion ?



How the growth of oil price ?

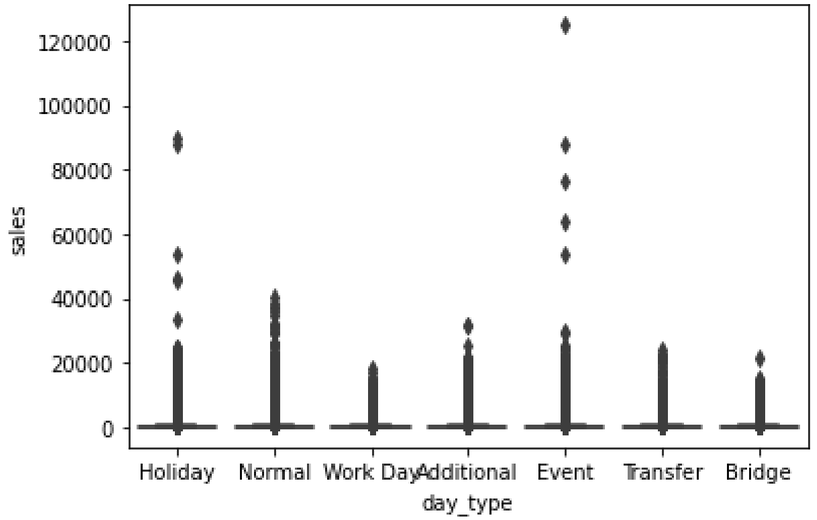


Deep-Dive Data Exploratory

2. How the correlation between sales, onpromotion and oilprice in monthly basis?

Observation :
the sum of sales, sum of onpromotion and average of oil price **in aggregat monthly basis has correlation** with each other. It's very different with standard EDA, that individual customer doesn't have significant correlation.

3. How the day type affect sales ?



Observation :

Customer usually buying in normal day (based on the result of Standard EDA), but in event day there's significant amount of buying product (outliers).

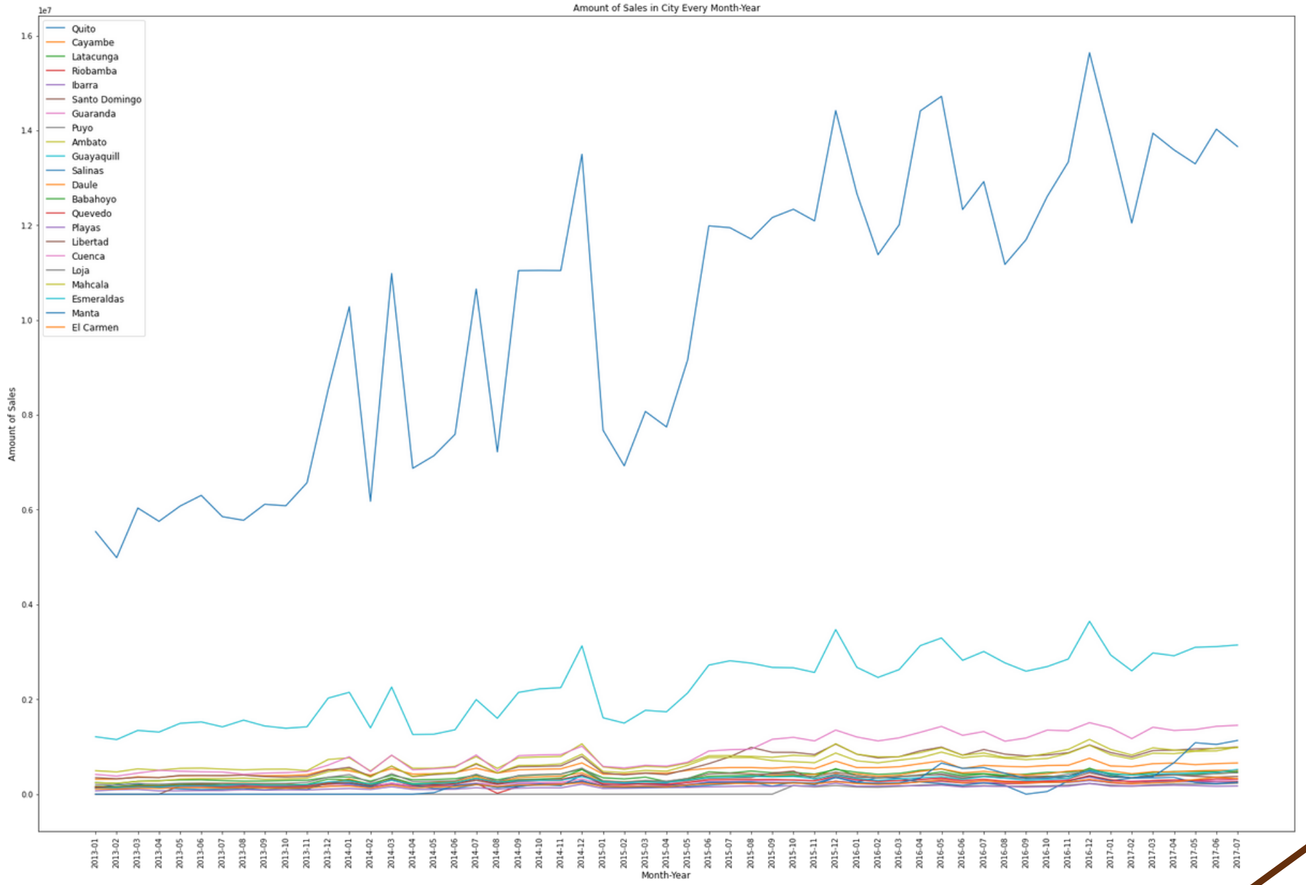
4. Who is the top 10 customer who bought product the largest ?

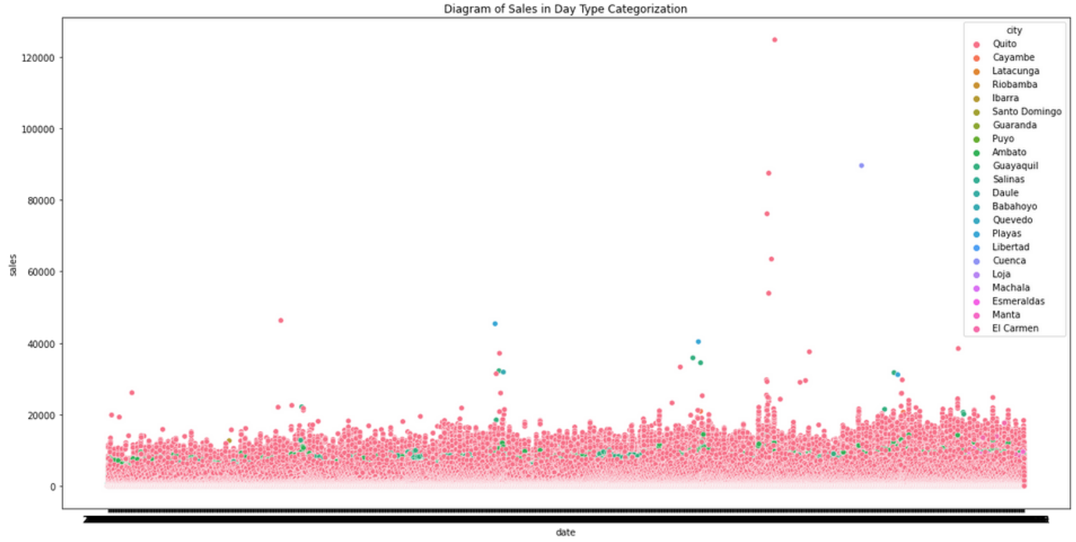
Observation :
Top 10 customer started with 48045 of product bought and maximum have bought 174877.032 products

	id	num_sales
2144154	2144154	174877.032
2163723	2163723	124717.000
2145045	2145045	107748.000
2445984	2445984	89576.360
2139699	2139699	76090.000
2153031	2153031	63434.000
2909844	2909844	52842.000
2144145	2144145	49224.000
2181576	2181576	48529.700
2909556	2909556	48045.000

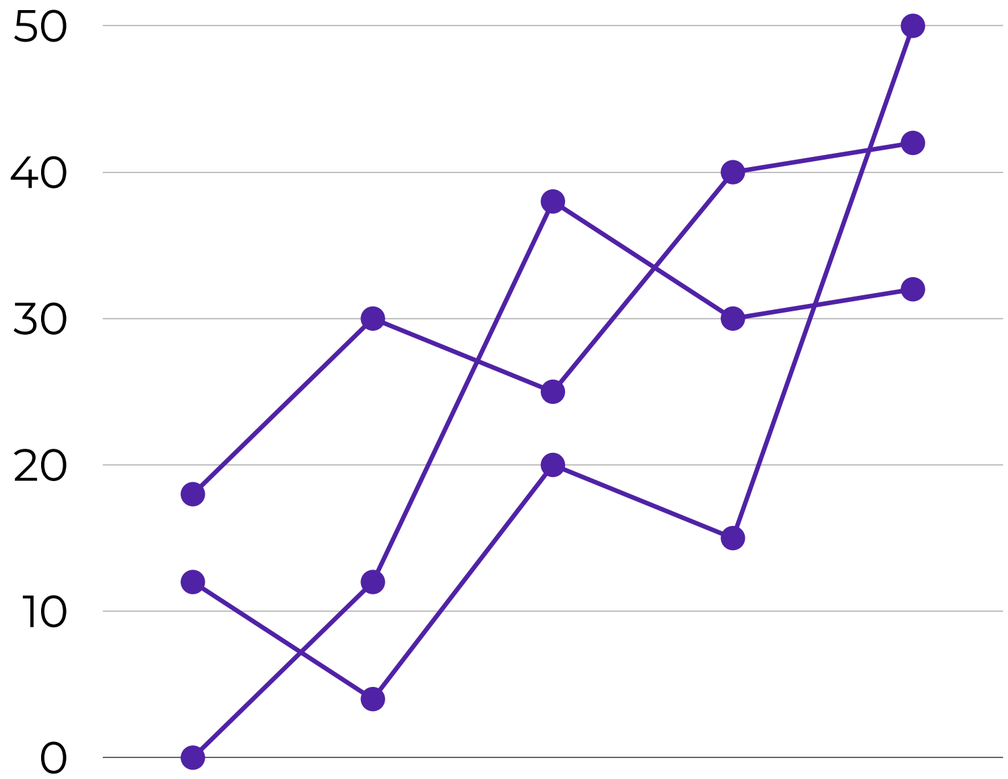
Deep-Dive Data Exploratory

5. How the growth of sales and product available in each city ?





Observation :
the sales is dominant in Quito, its inline with the result graph from standard EDA which the largest sales is in Quito., but every city in aggregate has stagnan growth



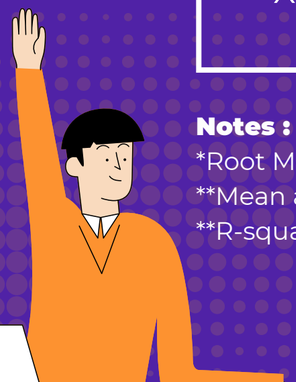
RESULT

Model Forecasting Sales

Model Forecasting

Result of Hyperparameter Tuning and applied best paramaters on Test Data

Model Forecasting	Parameters	RMSE*	MAE**	R2***	Result
Ridge Regression	alpha = 10	121.623	91.948	55,545	Declined
Lasso Regression	alpha = 0,0001	121.623	91.94	55,545%	Declined
Random Forest	max_depth=10, min_samples_leaf=4, n_estimators=15	53.265	22.224	91,472%	Declined (heavy computation)
XGB Boost	learning_rate = 0.5, max_depth = 10,n_estimators = 15	35.016	14.184	96,521%	**Accepted**



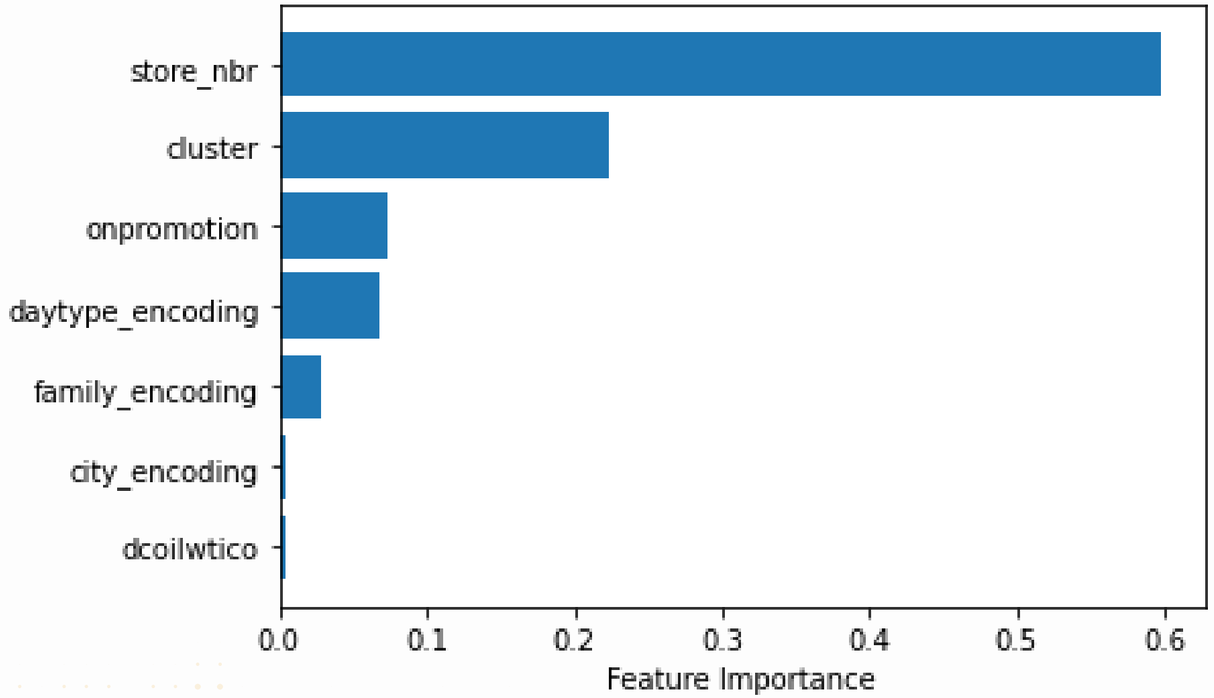
Notes :

*Root Mean Squared Error (RMSE) : the square root of the mean of the square of all of the error

**Mean absolute error (MAE) : Mean absolute error of actual data to predicted value

**R-squared (R2) : the percentage of variance in the dependent variable that can be explained by the independent variable

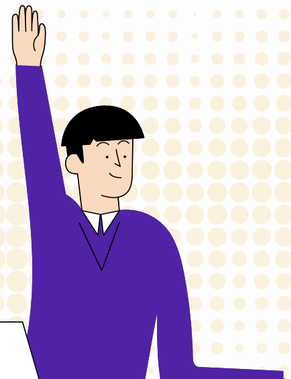
Feature Importance
XGB Boost



Observation :

From the result, we can know that :

- the loccation of the stores has significant impact to the sales amount
- Oil price has the lowest value of feature importance,
- There's no significant difference in RMSE, MAE and R-squared of train data and test data, which means XGB Boost model fit well!
- R-squared = 96, 521% shows that 96,521% of independent variables can explain the amount of sales (dependent variables). And, another 3,479% explain by others variables.



BUSINESS SOLUTIONS

Business Solution

1

Efficiency process of delivery/ allocation of product (supply chain) to reduce cost.

2

Concepting discount program.

- Discount program based on loyalty card from customers
- Discount program for any product regularly

3

Increase activity of digital marketing to increase awareness and soft marketing to engage with customers or prospective buyers

IMPACT ANALYSIS

The Weakness of Supply Chain without Forecasting Sales of Product :

- There's possibility of company to deliver stock of product more than once a month to every stores in each city in Ecuador (in this table, we assumes 3 times of delivery).
- Company can inaccurate in calculating book amount of stock by purchase order to supplier.
- Distribution of stock doesn;t represent the need of the stores.

The Impact of Supply Chain with Forecasting Sales of Product :

SAVE MORE THAN $3.652,42 or Rp 54.782.636,46

- Delivery can be once a month
- There's minimum risk of excess stock
- Distribution of stock can be fulfilled based on customer's purchasing power in each city
- *Special case* : if there's huge demonstration in Quito, there's no need of the company to fulfill the stock of product in other city in a range of date in a month

Delivery from - to	Distance (km)	Cost of oil (without forecasting supply chain)	Cost of oil (with forecasting supply chain)
Quito - Cayambe	68,8	$ 57,51	$ 19,17
Quito - Latacunga	108,1	$ 90,36	$ 30,12
Quito - Ibarra	112,9	$ 94,37	$ 31,46
Quito - Santo Domingo	152,9	$ 127,81	$ 42,60
Quito - Ambato	157,5	$ 131,65	$ 43,88
Quito - El Carmen	185,5	$ 155,06	$ 51,69
Quito - Riobamba	209,8	$ 175,37	$ 58,46
Quito - Guaranda	250	$ 208,97	$ 69,66
Quito - Puyo	251,7	$ 210,39	$ 70,13
Quito - Esmeraldas	317,3	$ 265,23	$ 88,41
Quito - Babahoyo	349,7	$ 292,31	$ 97,44
Quito - Daule	390,5	$ 326,41	$ 108,80
Quito - Manta	397,3	$ 332,10	$ 110,70
Quito - Guayaquil	421,5	$ 352,32	$ 117,44
Quito - Libertad	428	$ 357,76	$ 119,25
Quito - Cuenca	469	$ 392,03	$ 130,68
Quito - Machala	517,4	$ 432,49	$ 144,16
Quito - Playas	520,5	$ 435,08	$ 145,03
Quito - Salinas	568,2	$ 474,95	$ 158,32
Quito - Loja	677,7	$ 566,48	$ 188,83
Total (in $)		**$ 5.478,63**	**$ 1.826,21**
Total (in Rp (Kurs 18/09/2022))		**Rp 82.173.954,69**	**Rp 27.391.318,23**

Notes :

- Based on the latest solar oil price in Ecuador (28/06/22)= $1,9/ gallon (kompas.tv)
- 1 gallon = 4,546 Liter

References

Kaggle

https://www.kaggle.com/competitions/store-sales-time-series-forecasting/data

Kaggle

https://www.kaggle.com/code/moatazbellahahmed/stores-sales-prediction-eda

Thank You

Let's discuss & connect!

Github :
https://github.com/nymtw

Medium :
https://medium.com/@NiNyomanTriwahyuni

LinkedIn :
https://www.linkedin.com/in/ni-nyoman-triwahyuni/