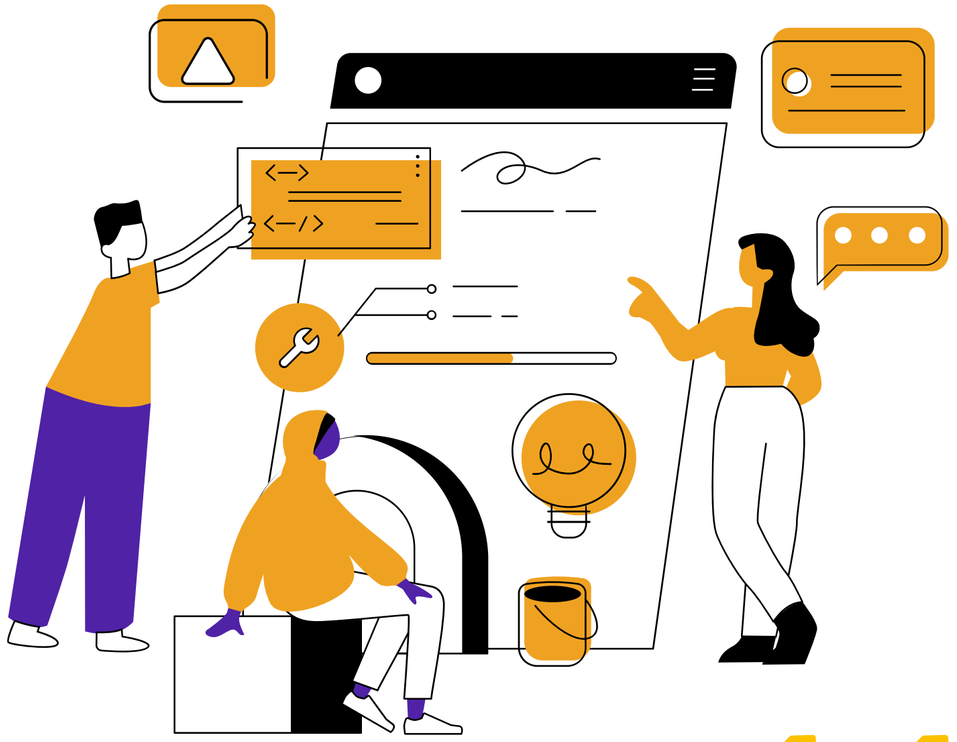
Forecasting

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

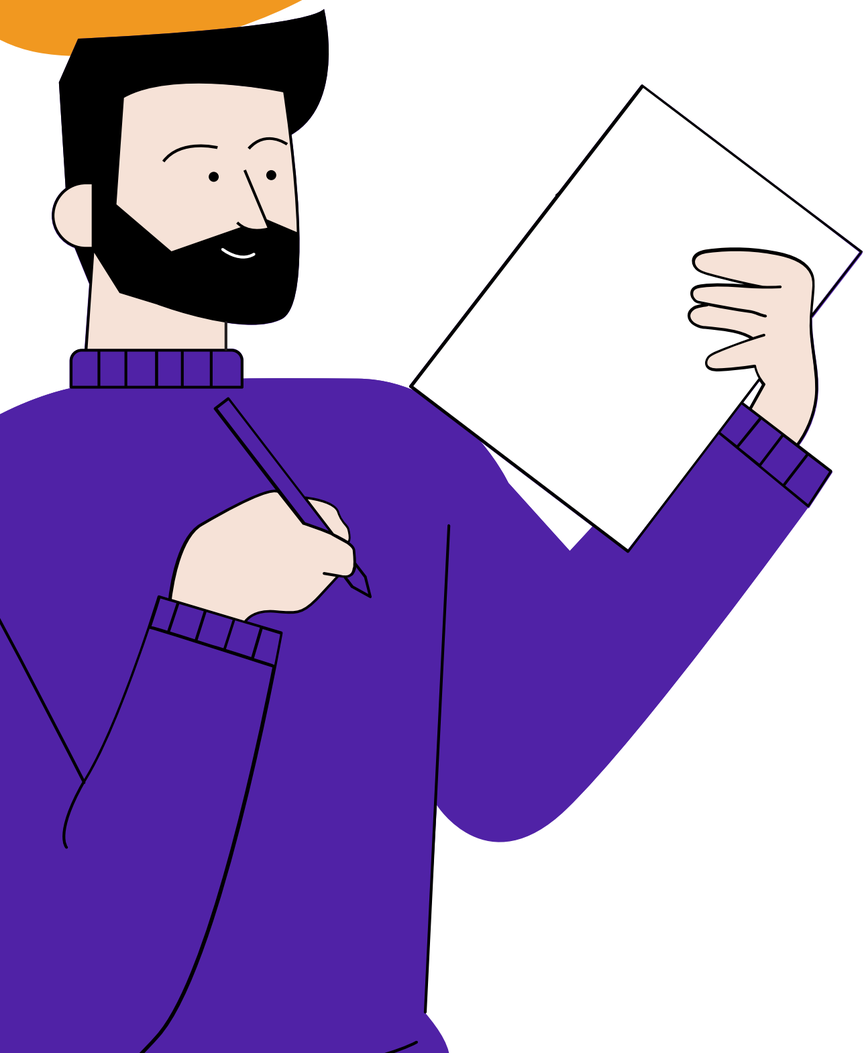
INTRODUCTION

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

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

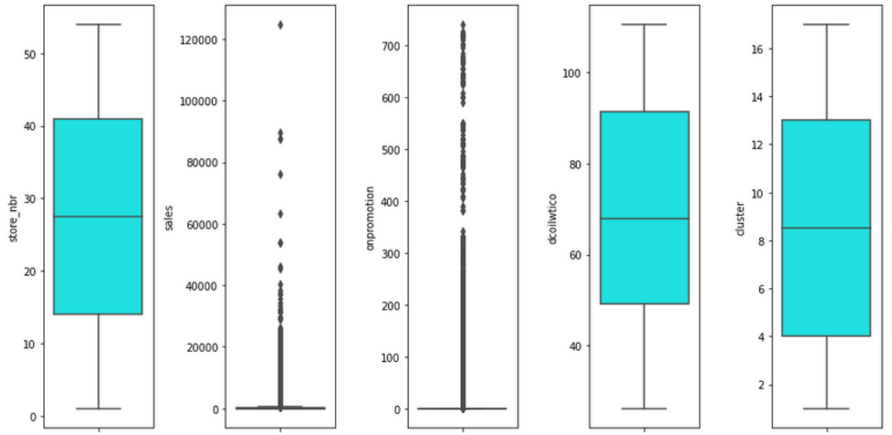
Numerical Variables

	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

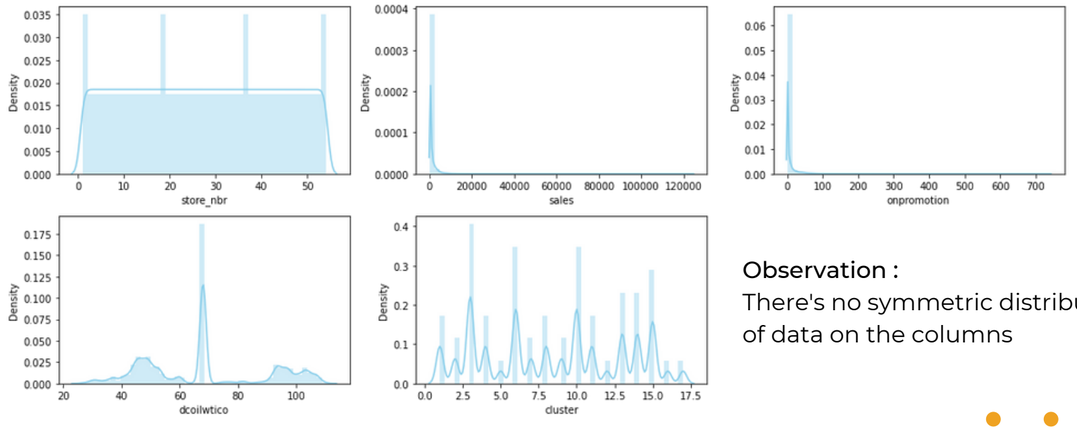
	family	city	state	day_type
count	3054348	3054348	3054348	3054348
unique	33	22	16	7
top	AUTOMOTIVE	Quito	Pichincha	Normal
freq	92556	1018116	1074678	2551824

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.

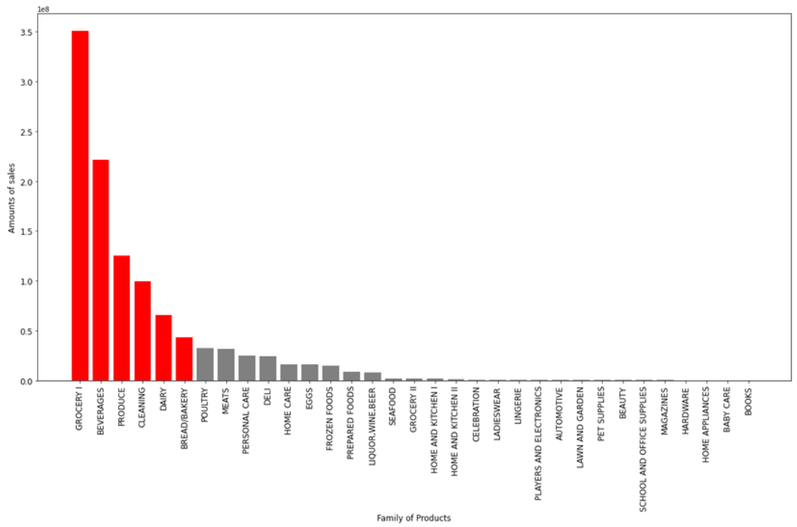


Observation :
There's no symmetric distribution of data on the columns

Standard Data Exploratory

Multivariate Analysis (Categorical Variables)

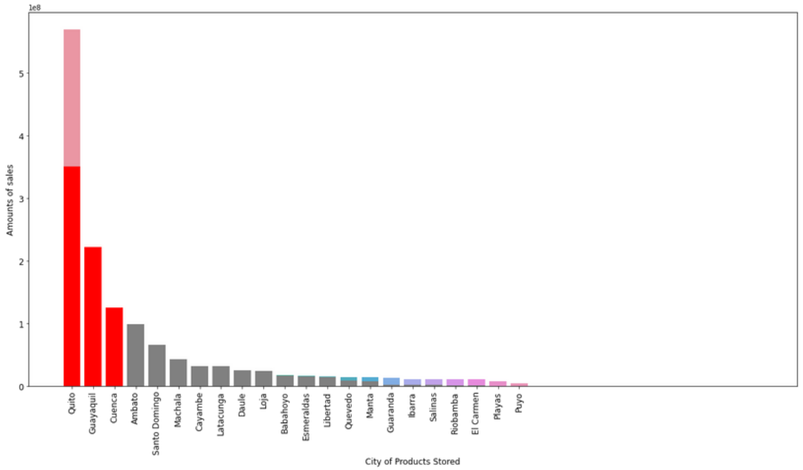
Amount of Sales Family of Product



Family of product which have sales more than average are

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

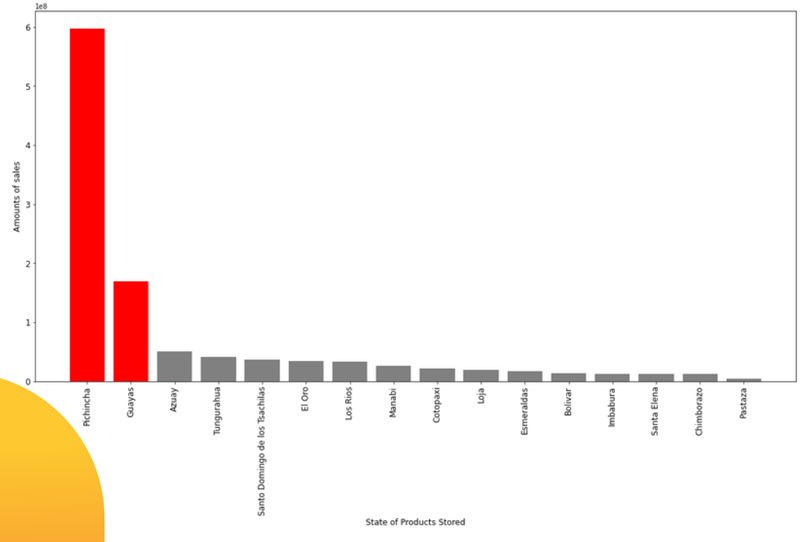
Amount of Sales in City



City which have sales more than average are :

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

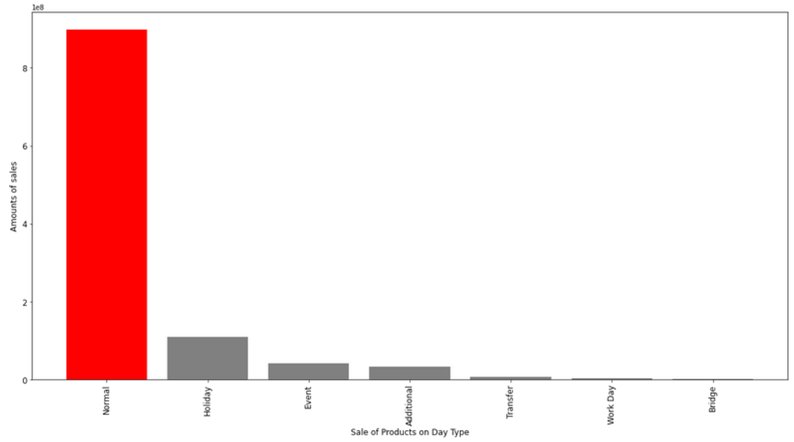
Amount of Sales in State



State which have sales more than average are :

- Pichincha
- Guayas

Amount of Sales in Each Day Type

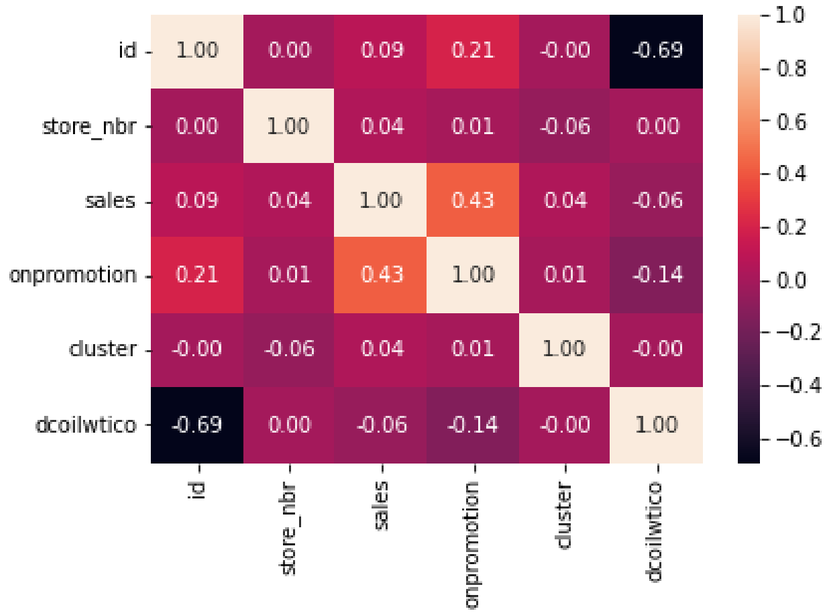


Day type which have sales more than average is on the "Normal" day.

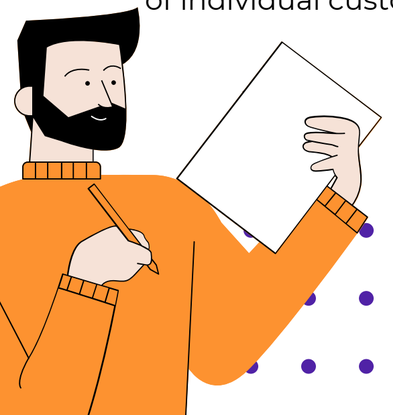
Standard Data Exploratory

Multivariate Analysis (Numerical Variables)

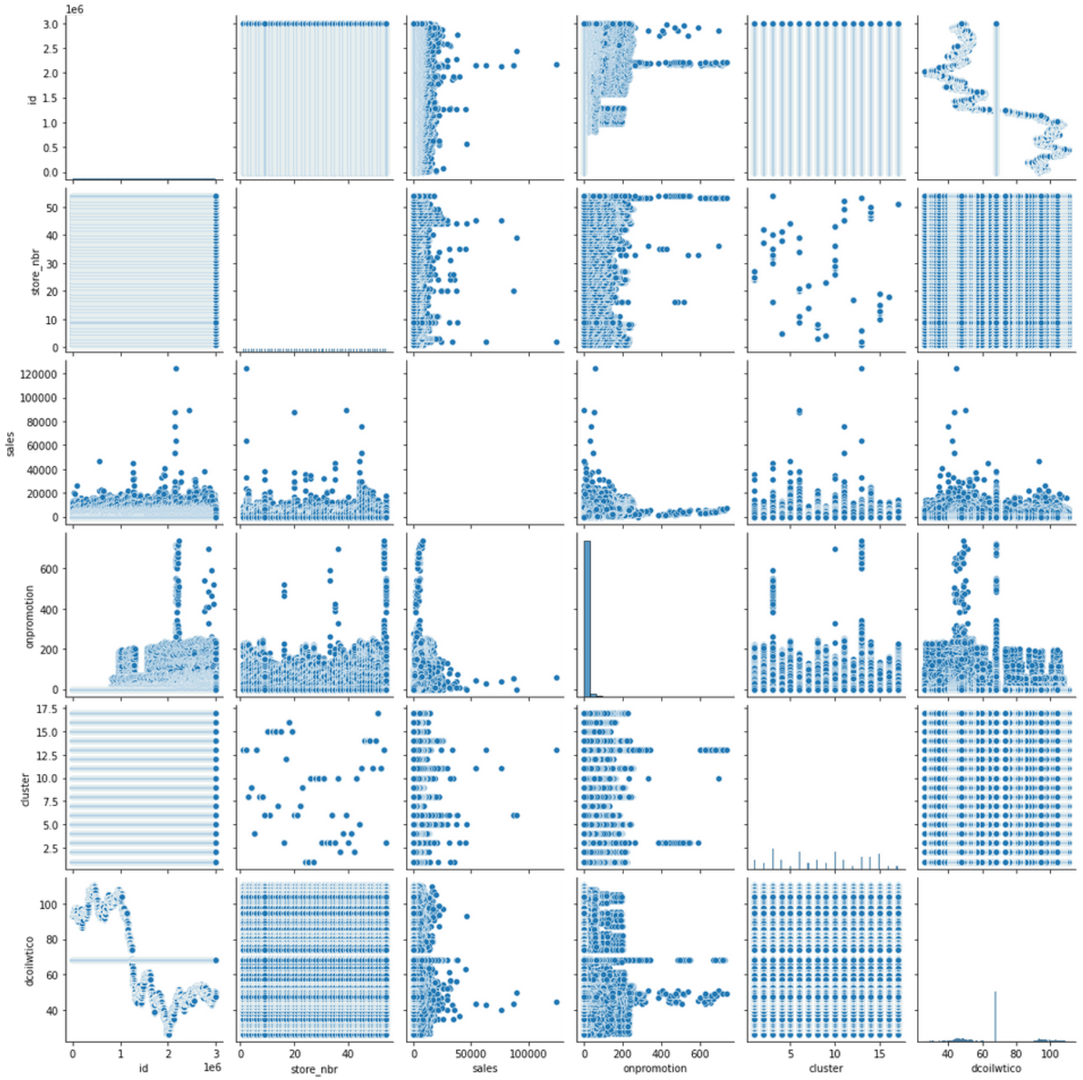
Heatmap Correlation



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



Paiplot of Numerical Variables



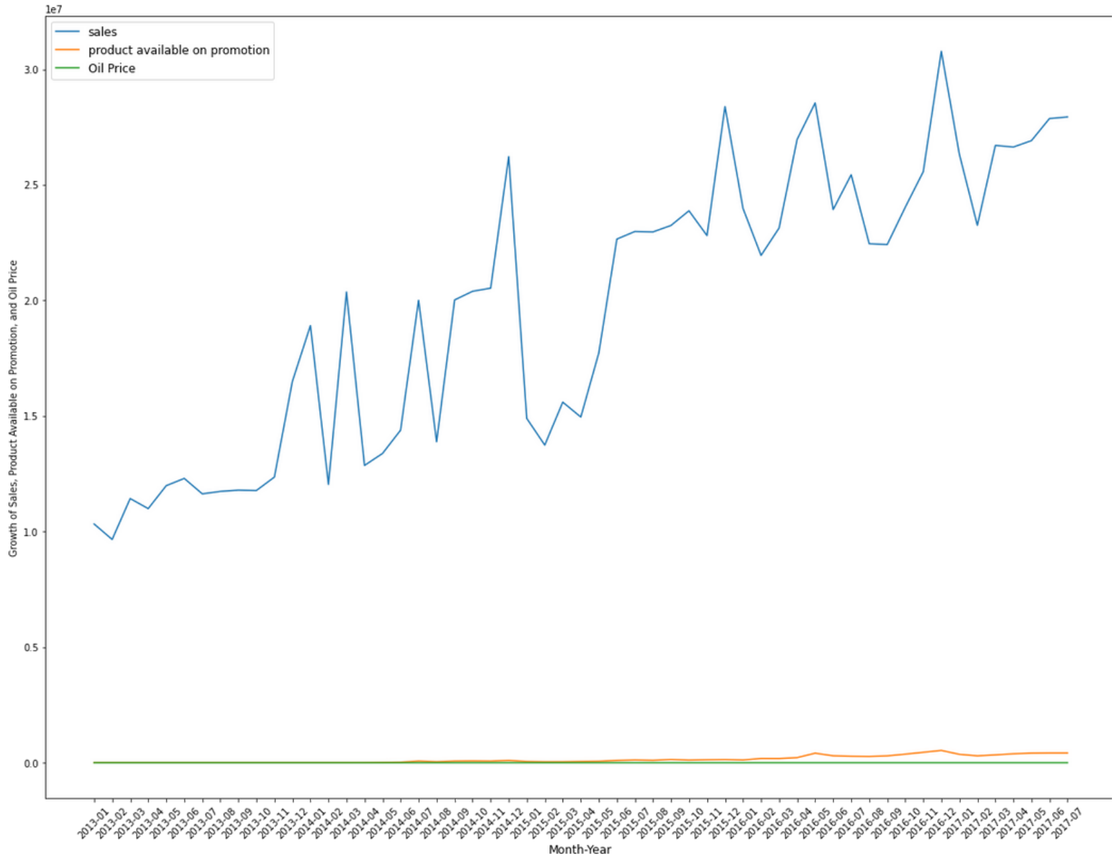
Observation :
From the graph we know that :sales and onpromotion evenly distribute data on store_nbr, cluster and dcoilwtico(oil price).

RESULT

2. Deep-Dive Data Explarotary

Deep-Dive Data Exploratory

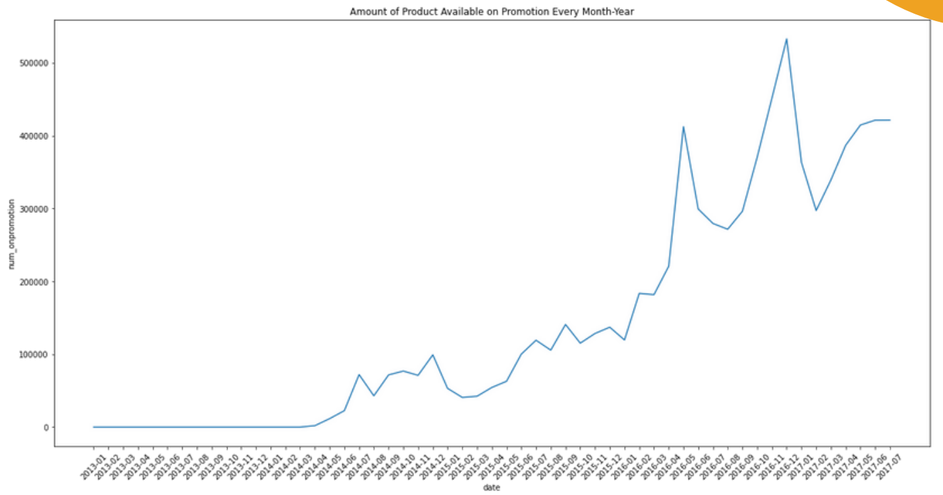
1. How the growth of sales, product available on promotion and oil price?



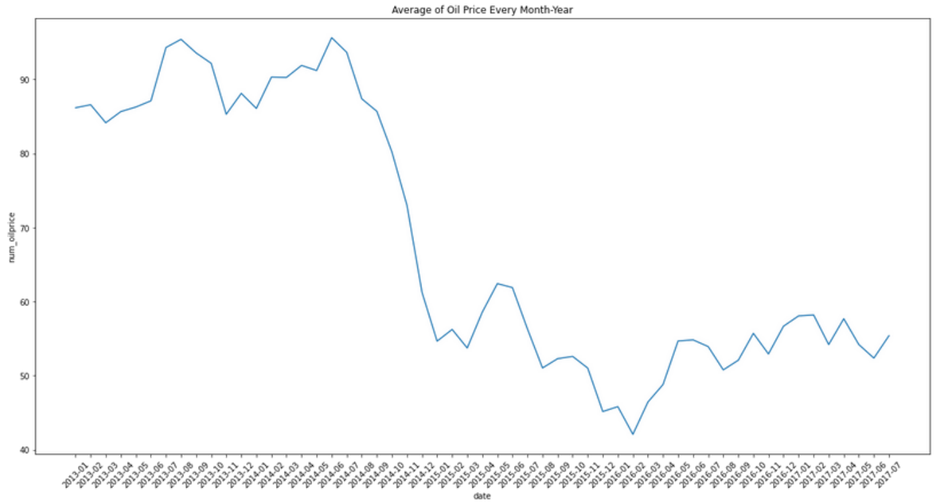
Observation :

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 ?

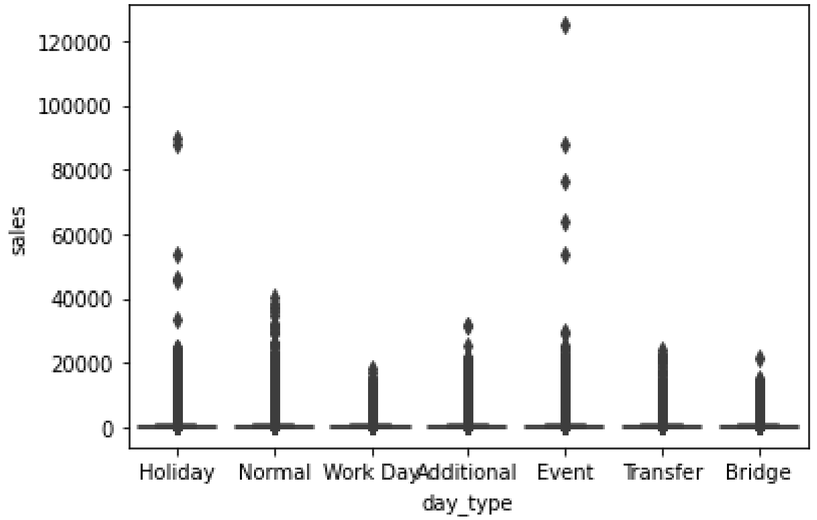


Deep-Dive Data Exploratory

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).

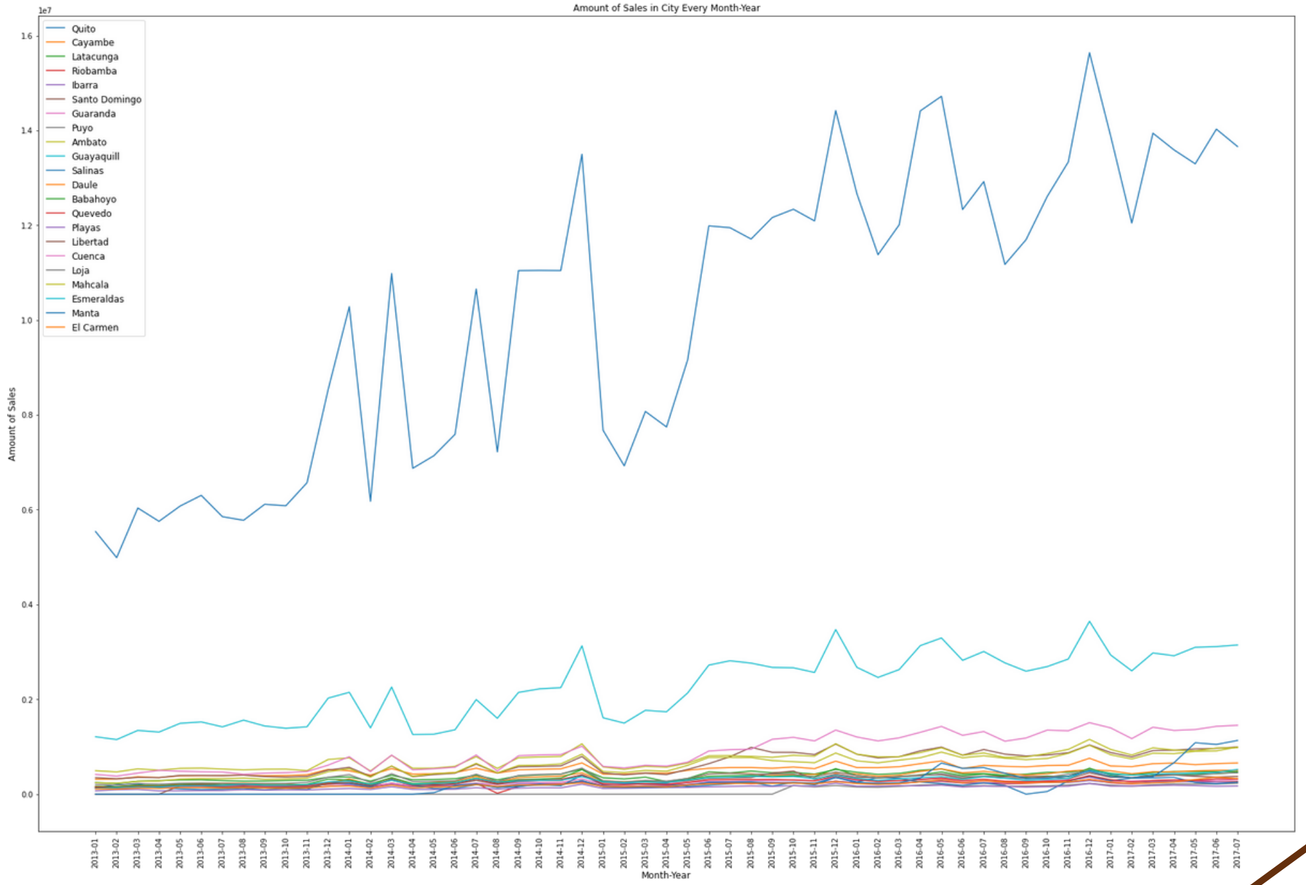
4. Who is the top 10 customer who bought product the largest ?

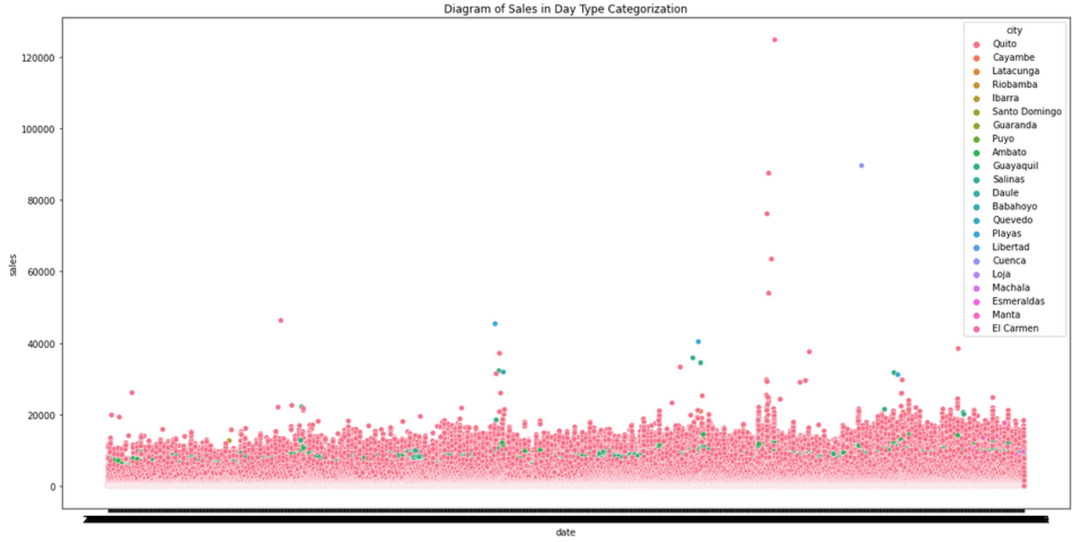
Observation :
Top 10 customer started with 48045 of product bought and maximum have bought 174877.032 products

	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

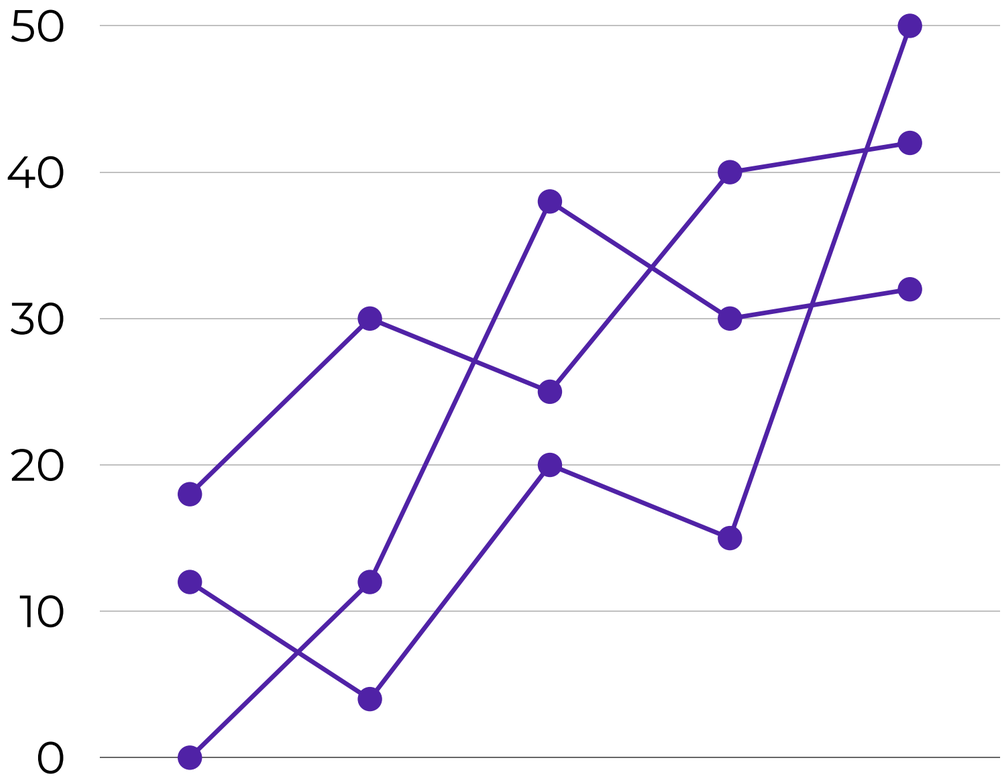
Deep-Dive Data Exploratory

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

3. Model Forecasting Sales

Model Forecasting

(Hyperparameter Tuning)

Ridge Regression

```
[113] # only show the most important columns
      retain_cols = ['params','mean_test_score','rank_test_score']
      cv_result = pd.DataFrame(ridge_reg_gridcv.cv_results_)
      cv_result[retain_cols]
```

	params	mean_test_score	rank_test_score
0	{'alpha': 1e-06}	-121.54747	8
1	{'alpha': 1e-05}	-121.54747	8
2	{'alpha': 0.0001}	-121.54747	8
3	{'alpha': 0.001}	-121.54747	7
4	{'alpha': 0.01}	-121.54747	6
5	{'alpha': 0.1}	-121.54747	5
6	{'alpha': 1}	-121.54747	4
7	{'alpha': 5}	-121.54747	3
8	{'alpha': 10}	-121.54747	2
9	{'alpha': 20}	-121.54747	1

```
[114] # the best model
      ridge_reg_gridcv.best_estimator_
```

```
Ridge(alpha=20, random_state=42)
```

Observation :
The best alpha (lambda) for ridge reggression is 20, because have the lowest mean test score.

Lasso Regression

```
# only show the most important columns
retain_cols1 = ['params','mean_test_score','rank_test_score']
cv_result1 = pd.DataFrame(lasso_reg_gridcv.cv_results_)
cv_result1[retain_cols]
```

	params	mean_test_score	rank_test_score
0	{'alpha': 1e-06}	-121.547470	3
1	{'alpha': 1e-05}	-121.547470	2
2	{'alpha': 0.0001}	-121.547470	1
3	{'alpha': 0.001}	-121.547470	4
4	{'alpha': 0.01}	-121.547470	5
5	{'alpha': 0.1}	-121.547549	6
6	{'alpha': 1}	-121.553956	7
7	{'alpha': 5}	-121.566425	8
8	{'alpha': 10}	-121.605389	9
9	{'alpha': 20}	-121.737236	10

```
[119] # the best model
      lasso_reg_gridcv.best_estimator_
```

```
Lasso(alpha=0.0001, random_state=42)
```

Observation :
TThe best alpha (lambda) for lasso regression is 0,0001, because have the lowest mean test score

Model Forecasting

(Training Any Model Data)

Ridge Regression

1.Train the data with top 3 lambda : 20, 10, 5

a. Using Data Train

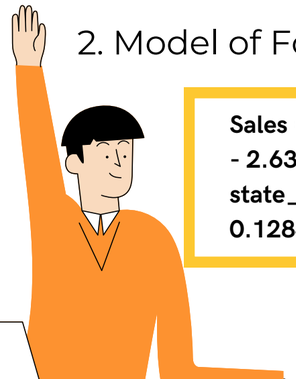
```
RMSE of Ridge regression model with alpha = 5 is 121.54693215829006
RMSE of Ridge regression model with alpha = 10 is 121.54693215829049
RMSE of Ridge regression model with alpha = 20 is 121.54693215829231
```

b. Using Data Validation

```
RMSE of Ridge regression model with alpha = 5 is 121.5348679051708
RMSE of Ridge regression model with alpha = 10 is 121.53486790517606
RMSE of Ridge regression model with alpha = 20 is 121.53486790519703
```

Conclusion :
Use alpha = 5 for modelling which have the lowest RMSE

2. Model of Forecasting



Sales = -9.606561 + 0.297776 city_encoding + 0.315201 store_nbr - 2.634238 family_encoding + 9.657971 onpromotion + 1.362759 state_encoding + 0.575253 Cluster - 1.622133 day_type encoding + 0.128448 dcoilwtico

Lasso Regression

1.Train the data with top 3 lambda : 0.0001, 1e-05, 1e-06

a. Using Data Train

```
RMSE of Lasso regression model with alpha = 1e-06 is 121.54693215828992
RMSE of Lasso regression model with alpha = 1e-05 is 121.54693215829072
RMSE of Lasso regression model with alpha = 0.0001 is 121.54693215829072
```

b. Using Data Validation

```
RMSE of Lassoregression model with alpha = 1e-06 is 121.53681453360063
RMSE of Lassoregression model with alpha = 1e-05 is 121.53681451820204
RMSE of Lassoregression model with alpha = 0.0001 is 121.53681451820204
```

Conclusion :
Because there's no significant difference in RMSE of alpha 0,0001 in train data and validate data, therefore will use lambda/alpha = 0,0001 which have the lowest RMSE

2. Model of Forecasting

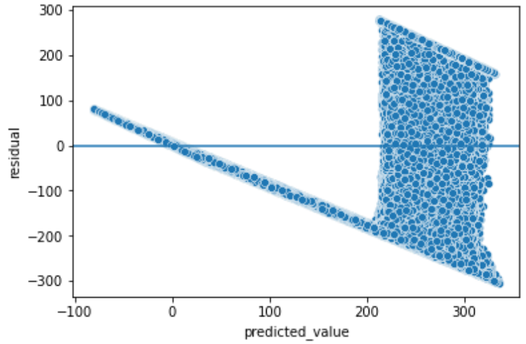
Sales = -9.607071 + 0.297775 city_encoding + 0.315201 store_nbr - 2.634237 family_encoding + 9.657970 onpromotion + 1.362754 state_encoding + 0.575249 Cluster - 1.621961 day_type encoding + 0.128447 dcoilwtico

Model Forecasting

(Model Evaluation)

Ridge Regression

1. Residual Plot



Observation :

1. linear relationship ? No there's consistent line
2. Variance constant ? No because there's variety of residuals even at the beginning there's consistent line
3. residual independent ? No because there's correlation

2 Evaluation

On train data :

1. RMSE = 121.54693215829006
2. MAE = 91.88065939540236
3. MAPE = 3.982593629834036e+16, which is not good. Because it's really have a high mean absolute error, that indicate the sales of product significantly different.

But from the result of test data :

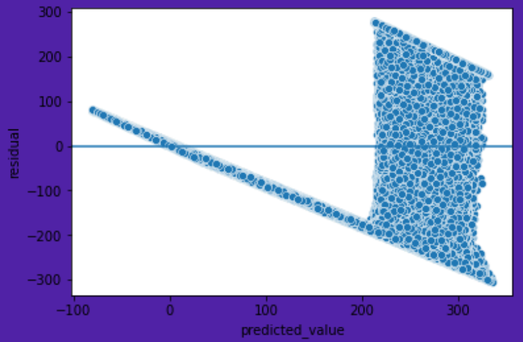
1. RMSE = 121.62305140000414
2. MAE = 91.94818620077154
3. MAPE = 3.972261653871982e+16, which is not good. Because it's really have a high mean absolute error, that indicate the sales of product significantly different.

3. Conclusion

- As the result of MAPE above we know that there's a acceptable for ridge model because the percentage of error no significant different from 3.982593629834036e+16(in train data) become 3.972261653871982e+16 (in test data)
- As the result of R-squared, 55,545% of variables can explain the value of sales (dependent variables). And, another 55,445% explain by others variables.
- **Underfitting Model, therefore need to add another variables.**

Lasso Regression

1. Residual Plot



Observation :

1. linear relationship ? Yes, there's consistent line
2. Variance constant ? No because there's variety of residuals even at the beginning there's consistent line
3. residual independent ? No because there's correlation

2. Evaluation

On train data :

1. RMSE = 121.54693215836905
2. MAE = 91.88065517100044
3. MAPE = 3.982591402460563e+16, which is not good. Because it's really have a high mean absolute error, that indicate the sales of product significantly different.

But from the result of test data :

1. RMSE = 121.62305145549408
2. MAE = 91.94818215247673
3. MAPE = 3.972259484836414e+16, which is not good. Because it's really have a high mean absolute error, that indicate the sales of product significantly different.

3. Conclusion

- As the result of MAPE above we know that there's a acceptable for ridge model because the percentage of error no significant different from 3.982591402460563e+16(in train data) become 3.972259484836414e+16 (in test data)
- As the result of R-squared, 55,545% of variables can explain the value of sales (dependent variables). And, another 55,445% explain by others variables.
- Underfitting Model, therefore need to add another variables.

BUSINESS SOLUTIONS

Business Solution

1

Analyze another factor that affect the customer to buy (ex: GDP)

2

Efficiency process of delivery/ allocation of product to reduce cost.

3

R&D of new product or enhance the innovation of existing product

4

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

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!