Predicting Store Sales

Nur Azizah

Linkedin: https://www.linkedin.com/in/nur-azizah-2a5233139/

Github: https://github.com/azizah717

Tableau Public: https://public.tableau.com/app/profile/nur.azizah5048

Summary

Data Introduction

Objective of Analysis

Result of Analysis

Business Solution

Data Introduction

- The data used for this analysis is Store Sales “Favorita” in Ecuador. In Ecuador, Favorita is one of the largest grocery chains.
- As an oil-dependent country, Ecuador's economy is highly sensitive to oil price fluctuations. Shipping products to grocery stores throughout the country affects inventory, and therefore sales.
- Source Data:

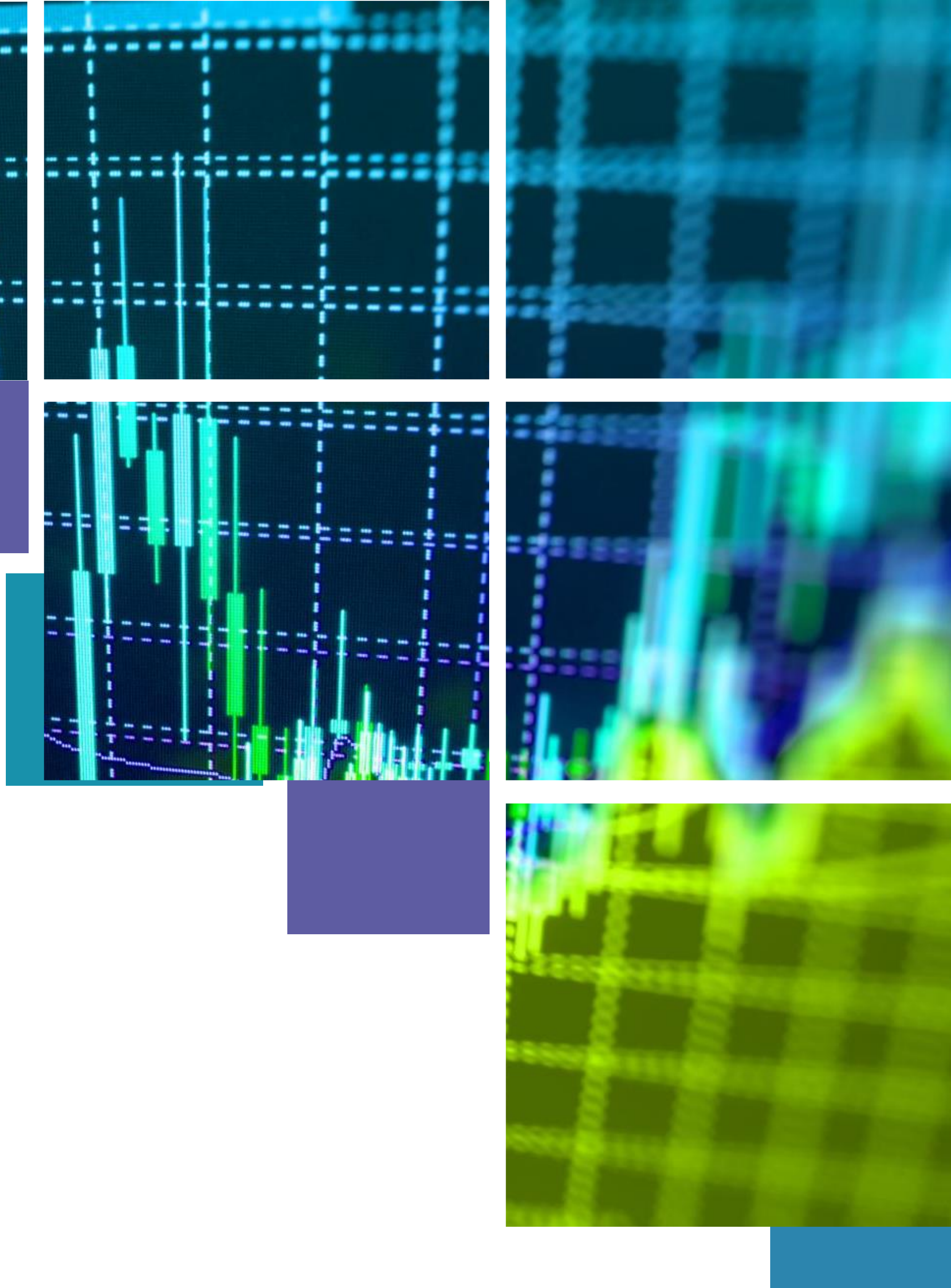
  https://www.kaggle.com/competitions/store-sales-time-series-forecasting/data

  (train, sample_submission, stores, oil, holidays_event)

Data Definition

- **Store_nbr** identifies the store at which the products are sold.
- **Family** identifies the type of product sold.
- **Sales** gives the total sales for a product family at a particular store at a given date
- **Onpromotion** gives the total number of items in a product family that were being promoted at a store at a given date.

- **Cluster** is a grouping of similar stores.
- **Dcoilwtico** oil price



Objective of Analysis

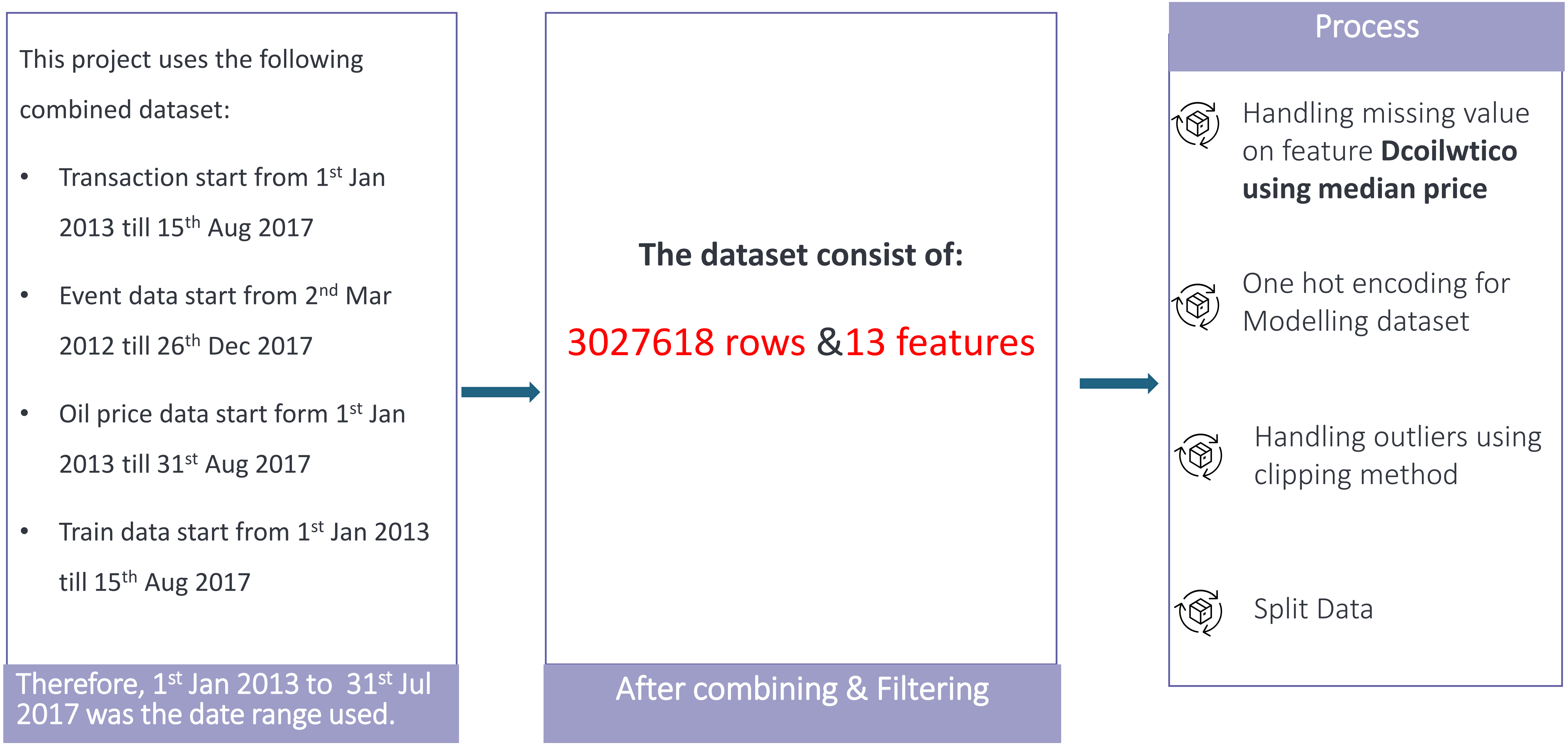
Exploratory Data Analysis

- Descriptive & Univariat Analysis
- Multivariate Analysis

Deep Dive Analysis

- How the growth of sales, product available on promotion and oilprice in monthly basis?
- How the correlation between sales, onpromotion, and oilprice is calculated in a monthly basis?
- When do sales have greatest impact?
- Who are the top biggest customers?
- Which cities are growing the most?

Dataset Information



This project uses the following combined dataset:

- Transaction start from 1st Jan 2013 till 15th Aug 2017
- Event data start from 2nd Mar 2012 till 26th Dec 2017
- Oil price data start form 1st Jan 2013 till 31st Aug 2017
- Train data start from 1st Jan 2013 till 15th Aug 2017

Therefore, 1st Jan 2013 to 31st Jul 2017 was the date range used.

The dataset consist of:

3027618 rows &13 features

After combining & Filtering

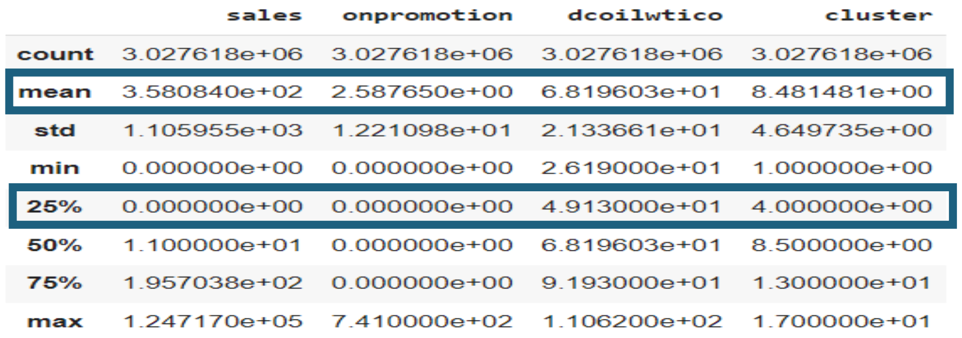
Process

- Handling missing value on feature **Dcoilwtico using median price**
- One hot encoding for Modelling dataset
- Handling outliers using clipping method
- Split Data

Descriptive & Univariate Analysis

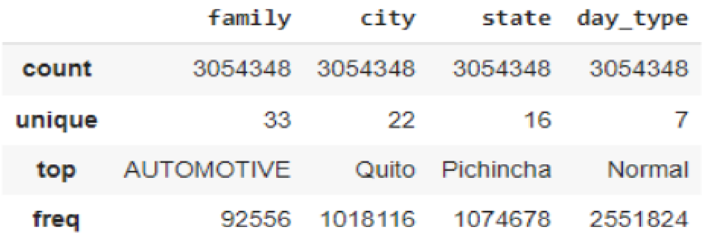
Descriptive Statistics

Numerical Features



	sales	onpromotion	dcoilwtico	cluster
count	3.027618e+06	3.027618e+06	3.027618e+06	3.027618e+06
mean	3.580840e+02	2.587650e+00	6.819603e+01	8.481481e+00
std	1.105955e+03	1.221098e+01	2.133661e+01	4.649735e+00
min	0.000000e+00	0.000000e+00	2.619000e+01	1.000000e+00
25%	0.000000e+00	0.000000e+00	4.913000e+01	4.000000e+00
50%	1.100000e+01	0.000000e+00	6.819603e+01	8.500000e+00
75%	1.957038e+02	0.000000e+00	9.193000e+01	1.300000e+01
max	1.247170e+05	7.410000e+02	1.106200e+02	1.700000e+01

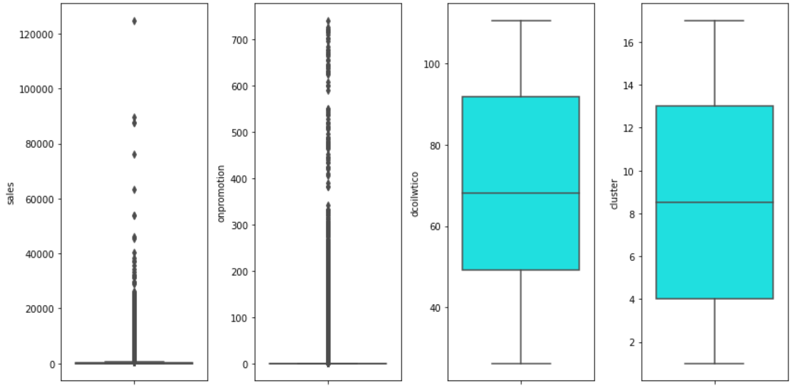
Categorical Features

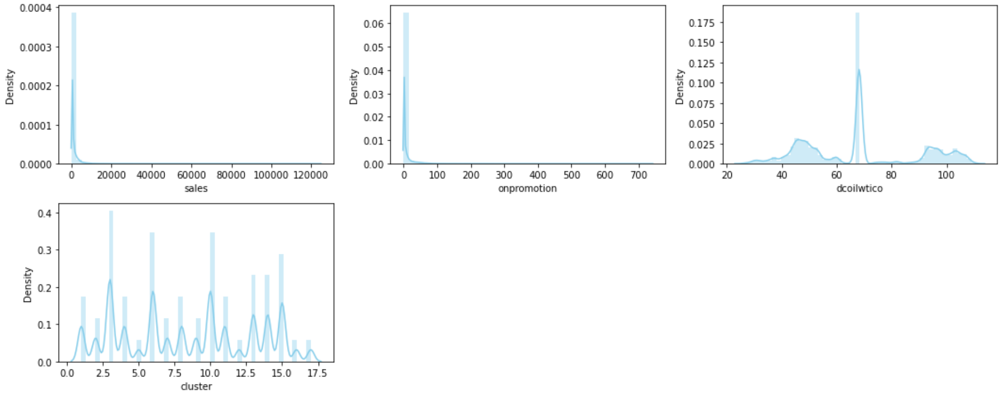


	family	city	state	day_type
count	3054348	3054348	3054348	3054348
unique	33	22	16	7
top	AUTOMOTIVE	Quito	Pichincha	Normal
freq	92556	1018116	1074678	2551824

- Sales and onpromotion not indicating a symmetrical distribution by looking at the mean-median difference
- Dcoilwtico and cluster indicating a symmetrical distribution
- The values of family, city, state, and day_type are unique

Univariate Analysis

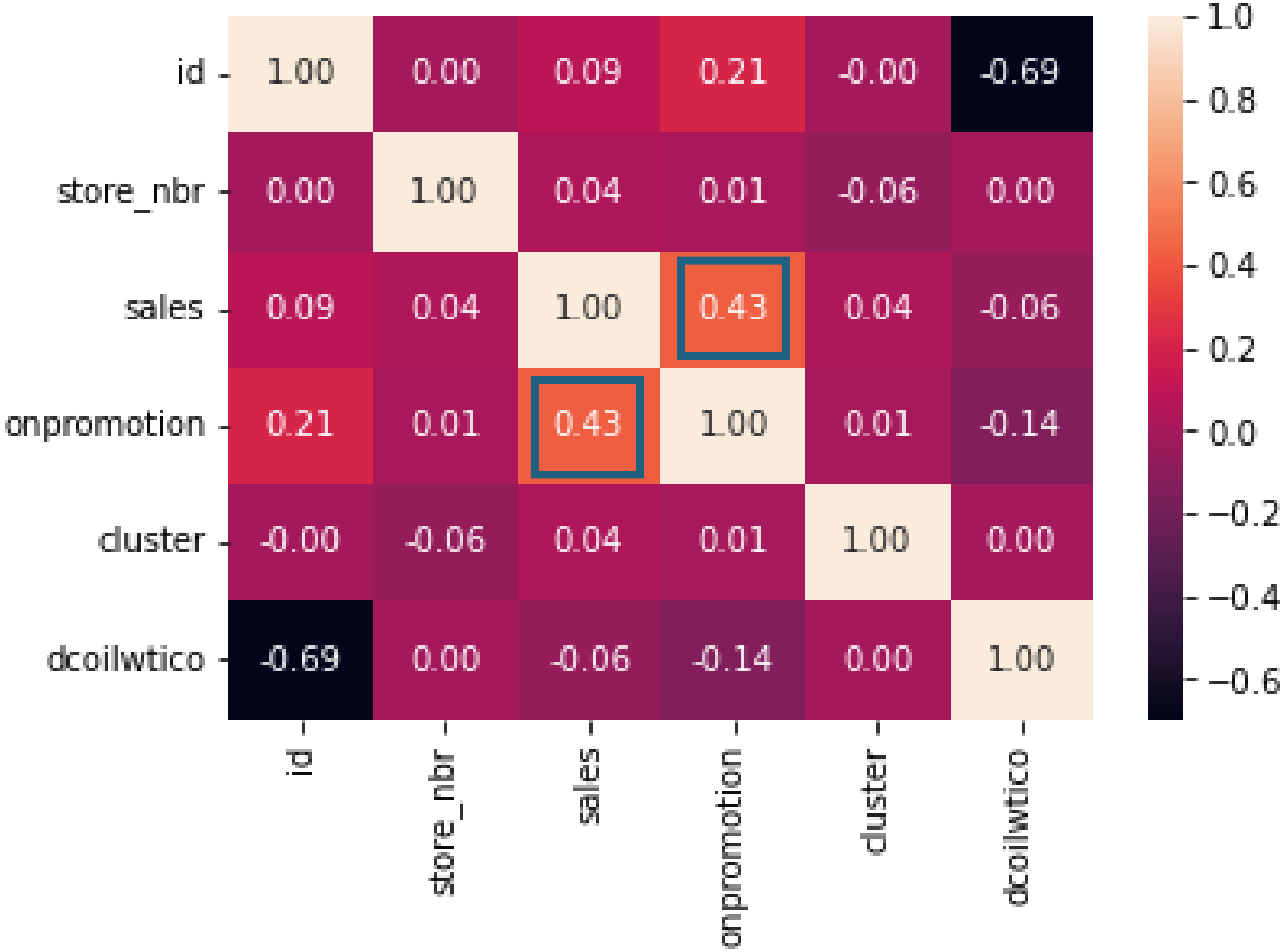




Except for sales and onpromotion, there are no outliers in store_nbr, dcoilwtico, or cluster. The data distribution of sales and onpromotion is not symmetric and there are many outliers.

Multivariate Analysis

Heatmap Correlation

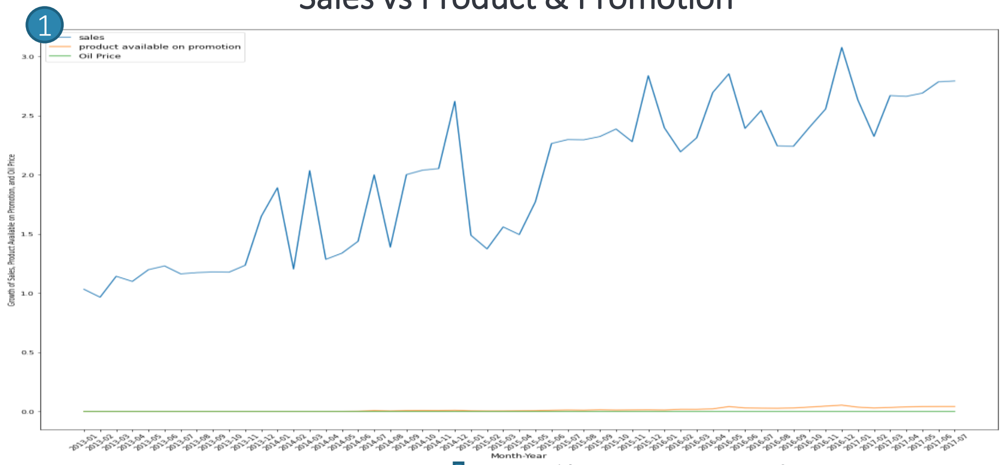


The largest correlation plot for sales and promotion characteristics is shown with probability = 0.43. This shows that both are positively related and that sales increase with more products advertised.

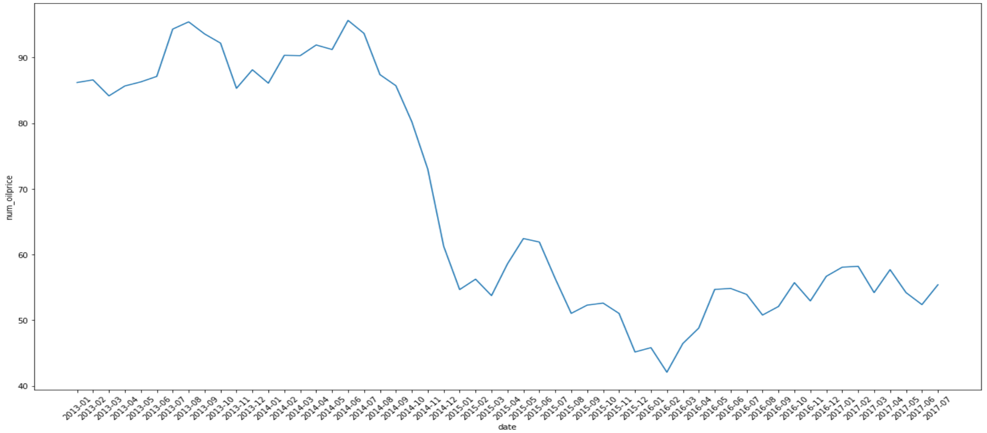
Result of Analysis

How the growth of sales, product available on promotion and oilprice in monthly basis?

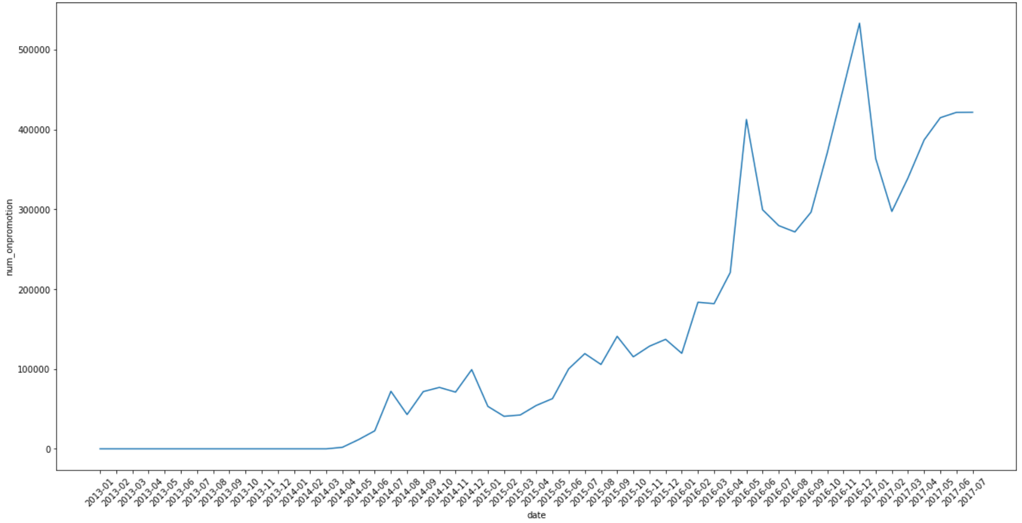
Sales vs Product & Promotion



Detail from oil price.



Detail from **the total number of items in a product family that were being promoted at a store at a given date.**



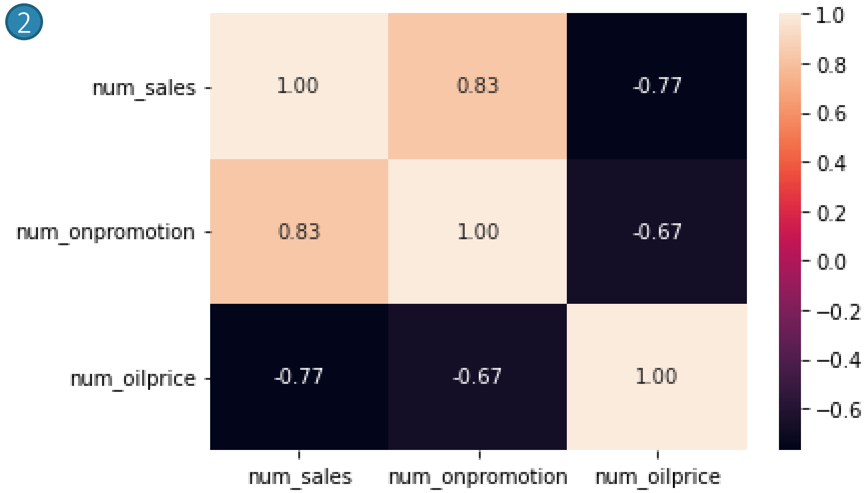
Conclusion:

1 Due to promotions and oil prices there is a large difference in sales volume compared to solid product available but overall sales increased from January 2013 to July 2017. This coincides with the increase in production available for promotion and the drop in oil prices that really characterizes Ecuador that sensitive to oil prices.

Result of Analysis

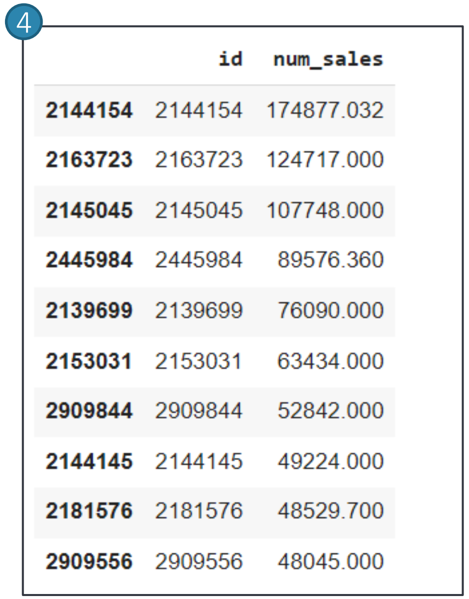
How the correlation between sales, onpromotion, and oilprice is calculated in a monthly basis?

2



Who are the top biggest customers?

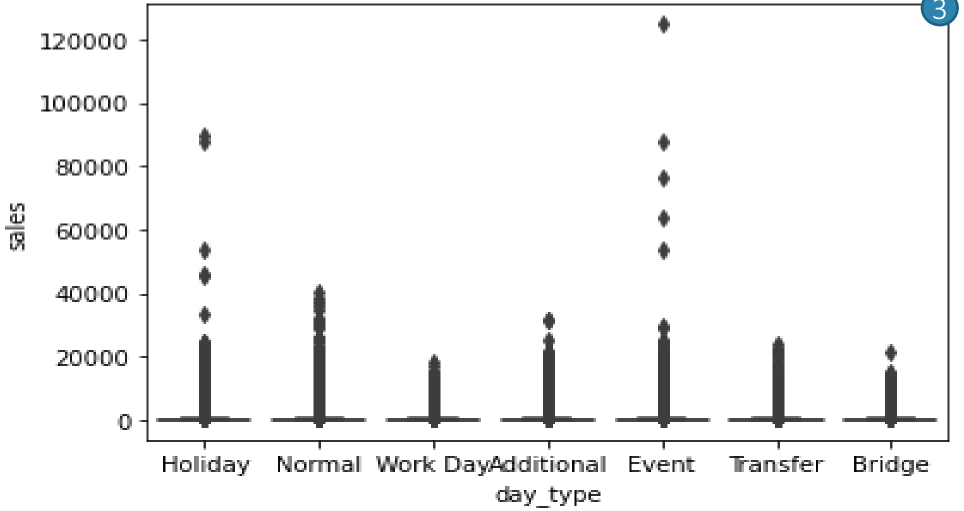
4



	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

When do sales have greatest impact?

3



Conclusion:

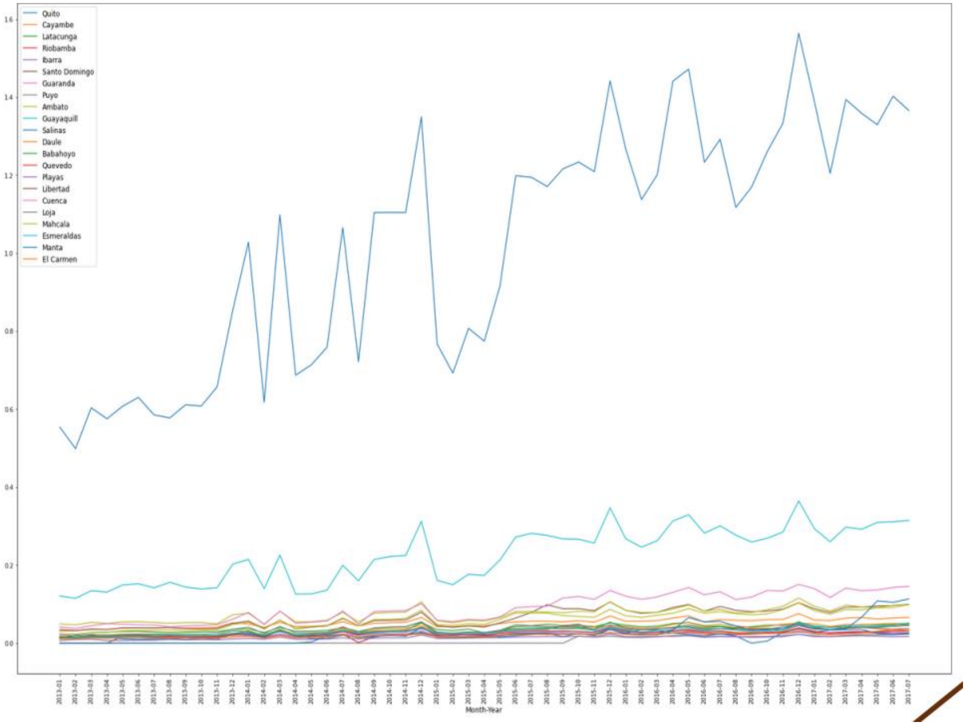
2 Promotion is positively correlated to sales, while oil prices are negatively correlated to sales with a strong correlation.

3 Sales with the highest consistency figures are on normal days. Whereas at events & holidays there are outliers where this is possible because of customer behavior

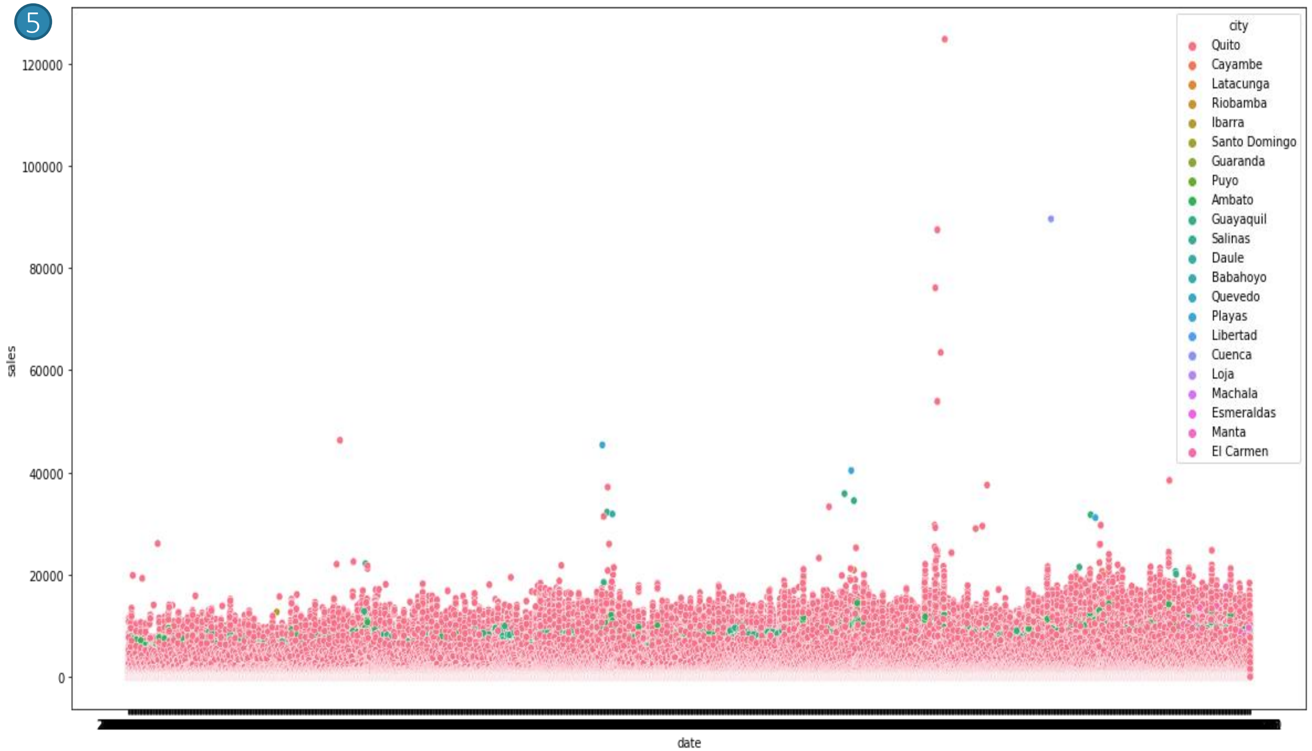
4 The customer with that ID is the customer with the highest number of purchases

Result of Analysis

Which cities are growing the most?



5



5 According to the result graph from a conventional EDA, Quito has the highest sales, hence Quito is clearly the market leader in terms of sales.

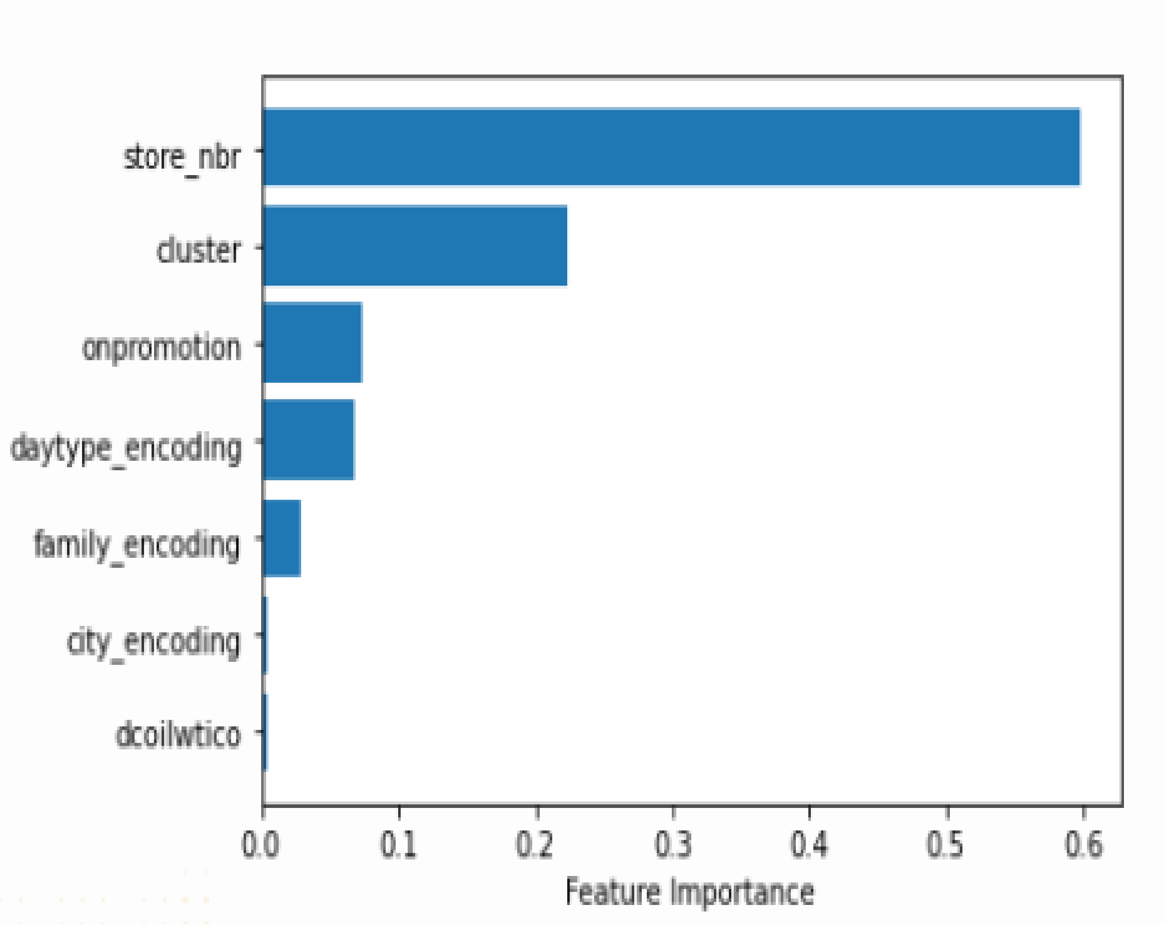
Model Prediction

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=4, n_estimator=15	53.265	22.224	91,472%	Declined
XGBoost	Learning_rate=0.5, max_depth=10, n_estimatord=15	35.016	14.184	96,521%	Accepted

XGBoost

Features Importance



- The location of the stores has significant impact to the sales amount
- The RMSE, MAE, and R-squared of the train and test data are not significantly different, indicating that the XGB Boost model fits the data well.
- R-squared = 96,521% demonstrates that 96,521% of independent variables can account for the volume of sales (dependent variables). Additionally, 3,479% more are explained by other factors.

Business Solution

1. To cut costs, the supply chain's delivery and allocation processes must be efficient.

2. Creating a discount scheme.
   - Discount scheme based on client loyalty cards
   - A recurring discount program for any goods

3. Increase soft marketing and digital marketing efforts to connect with customers and potential purchasers.

THANK YOU