



STORE SALES - TIME SERIES FORECASTING

USE MACHINE LEARNING TO PREDICT GROCERY SALES

Diego Alfonso Ramirez Montes

A01707596

Jose Angel Garcia Lopez

A01275108

Javier Suarez Duran

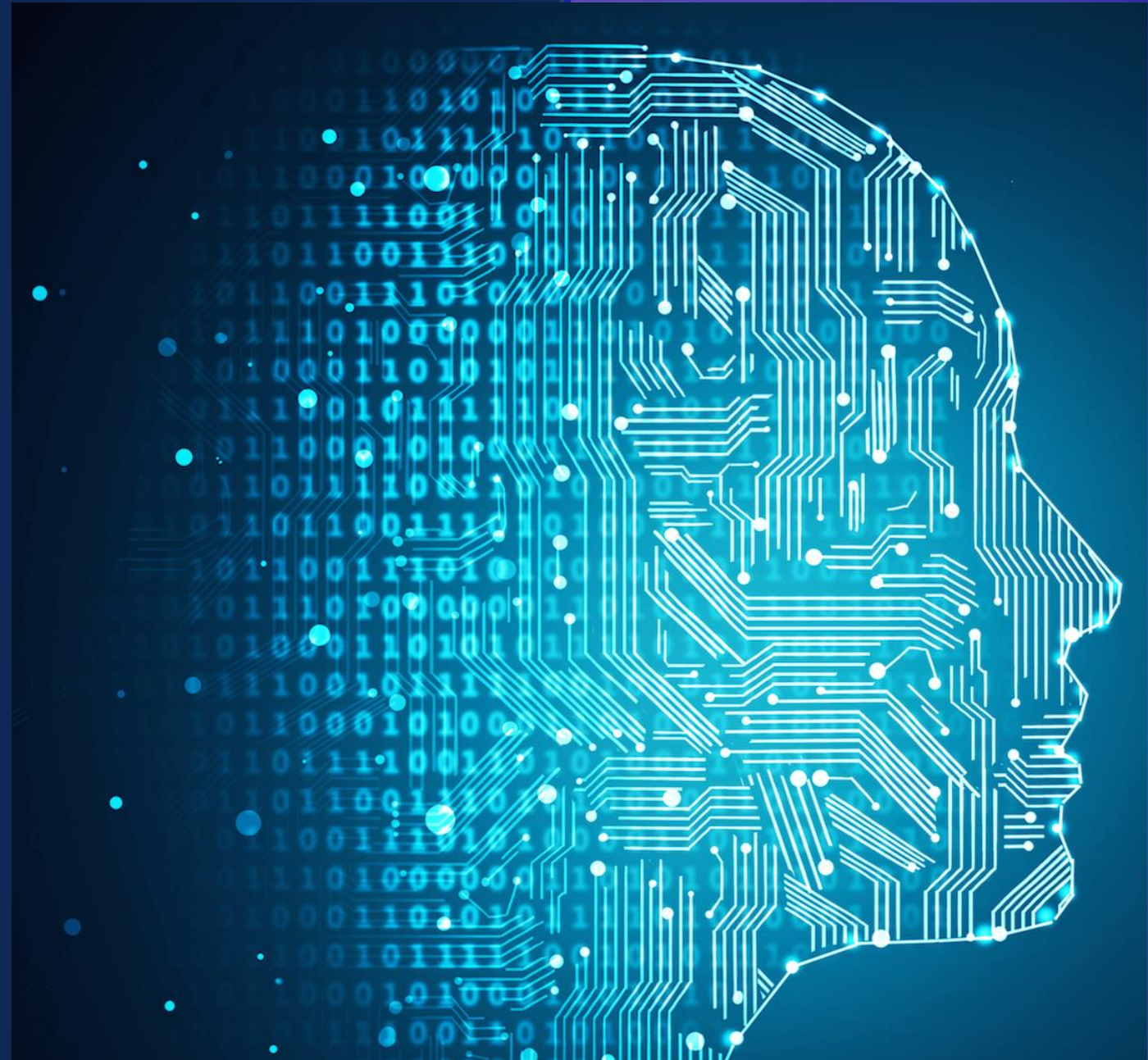
A01707380

Emiliano Mendoza Nieto

A01706083

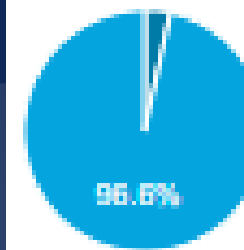
CONTENTS

- INTRODUCCIÓN
- ETL
- ANALISIS MODELOS
- METRICAS
- RESULTADOS





	date	type	locale	locale_name	description	transferred
345	2017-12-22	Additional	National	Ecuador	Navidad-3	False
346	2017-12-23	Additional	National	Ecuador	Navidad-2	False
347	2017-12-24	Additional	National	Ecuador	Navidad-1	False
348	2017-12-25	Holiday	National	Ecuador	Navidad	False
349	2017-12-26	Additional	National	Ecuador	Navidad+1	False

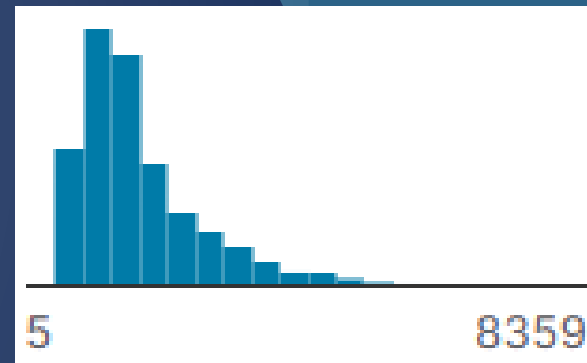
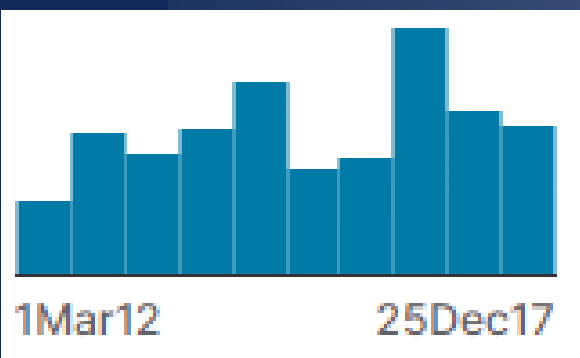


true
12 3%

false
338 97%

INTRODUCTION

- Goal of Competition
- Description Dataset



	id	date	store_nbr	family	sales	onpromotion
3000883	3000883	2017-08-15	9	POULTRY	438.133	0
3000884	3000884	2017-08-15	9	PREPARED FOODS	154.553	1
3000885	3000885	2017-08-15	9	PRODUCE	2419.729	148
3000886	3000886	2017-08-15	9	SCHOOL AND OFFICE SUPPLIES	121.000	8
3000887	3000887	2017-08-15	9	SEAFOOD	16.000	0

holidays_events.csv (22.31 kB)

Detail Compact Column

📅 date	⬆ type	⬆ locale	⬆ locale_name	⬆ description	✓ transferred
2012-03-02	Holiday	Local	Manta	Fundacion de Manta	False
2012-04-01	Holiday	Regional	Cotopaxi	Provincializaci on de Cotopaxi	False
2012-04-12	Holiday	Local	Cuenca	Fundacion de Cuenca	False
2012-04-14	Holiday	Local	Libertad	Cantonizacion de Libertad	False

stores.csv (1.39 kB)

Detail Compact Column

# store_nbr	⬆ city	⬆ state	⬆ type	# cluster
1	Quito	Pichincha	D	13
2	Quito	Pichincha	D	13
3	Quito	Pichincha	D	8
4	Quito	Pichincha	D	9
5	Santo Domingo	Santo Domingo de los Tsachilas	D	4

oil.csv (20.58 kB)

Detail Compact Column

📅 date	# dcoilwtico
2013-01-01	
2013-01-02	93.14
2013-01-03	92.97
2013-01-04	93.12
2013-01-07	93.2

train.csv (121.8 MB)

Detail Compact Column

 id 	 date 	 store_nbr 	 family 	 sales 	 onpromoti... 
0	2013-01-01	1	AUTOMOTIVE	0.0	0
1	2013-01-01	1	BABY CARE	0.0	0
2	2013-01-01	1	BEAUTY	0.0	0
3	2013-01-01	1	BEVERAGES	0.0	0



test.csv (1.02 MB)

Detail Compact Column

 id 	 date 	 store_nbr 	 family 	 onpromoti... 
3000888	2017-08-16	1	AUTOMOTIVE	0
3000889	2017-08-16	1	BABY CARE	0
3000890	2017-08-16	1	BEAUTY	2
3000891	2017-08-16	1	BEVERAGES	20

transactions.csv (1.55 MB)

Detail Compact Column

 date 	 store_nbr 	 transactions 
2013-01-01	25	770
2013-01-02	1	2111
2013-01-02	2	2358
2013-01-02	3	3487
2013-01-02	4	1922

EXTRACT TRANSFORM LOAD



CRITERIOS A SEGUIR

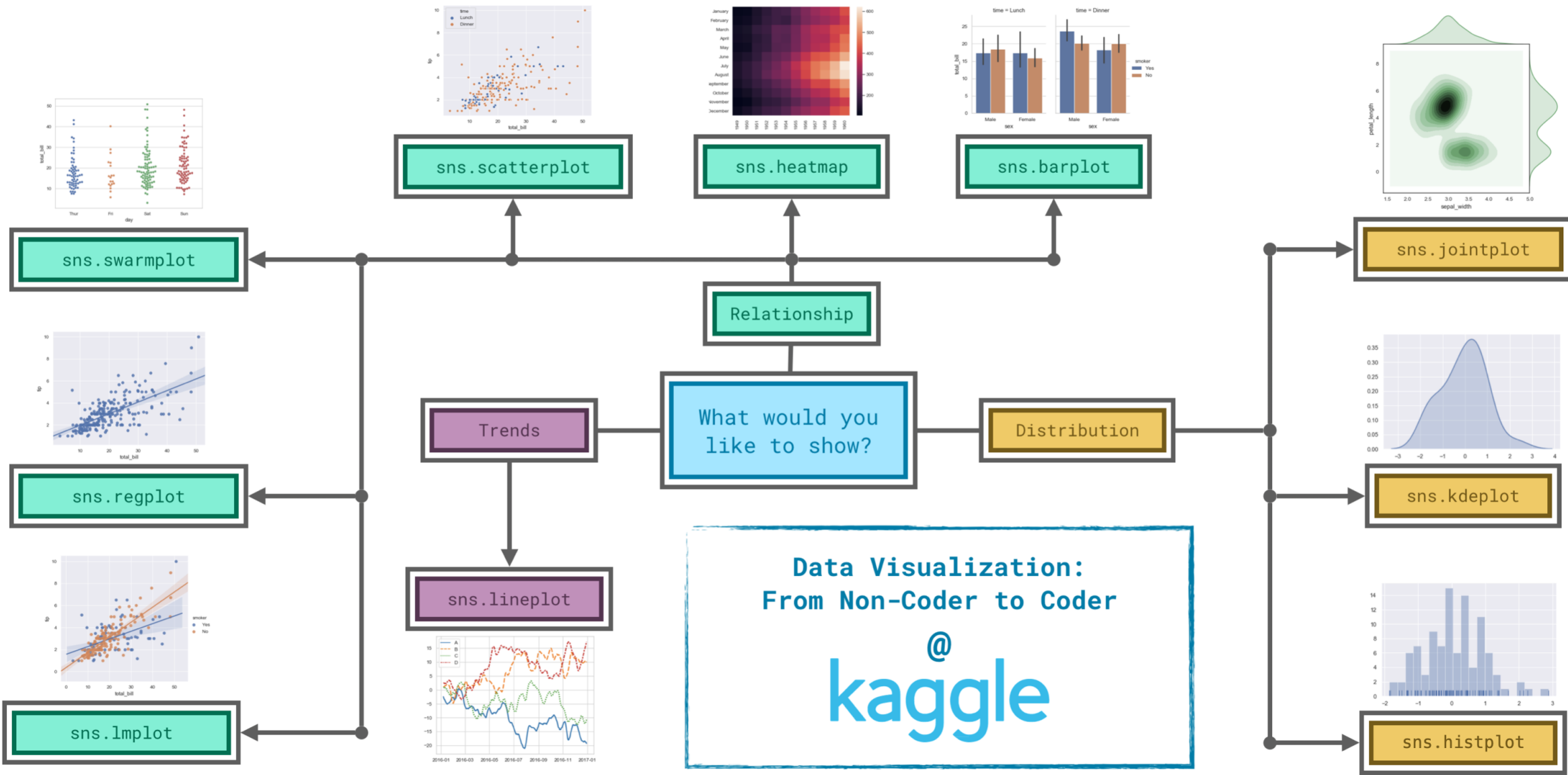
- Lo realizado en *train*, aplicaria tambien a *test*.
- Elegimos que variables parecen afectar mas a la variable a predecir "*sales*"
- Tomamos los datos que cambian atraves del tiempo
- Eliminar columnas con grandes cantidades datos faltantes
- Imputar datos de las columnas con pocos datos faltantes



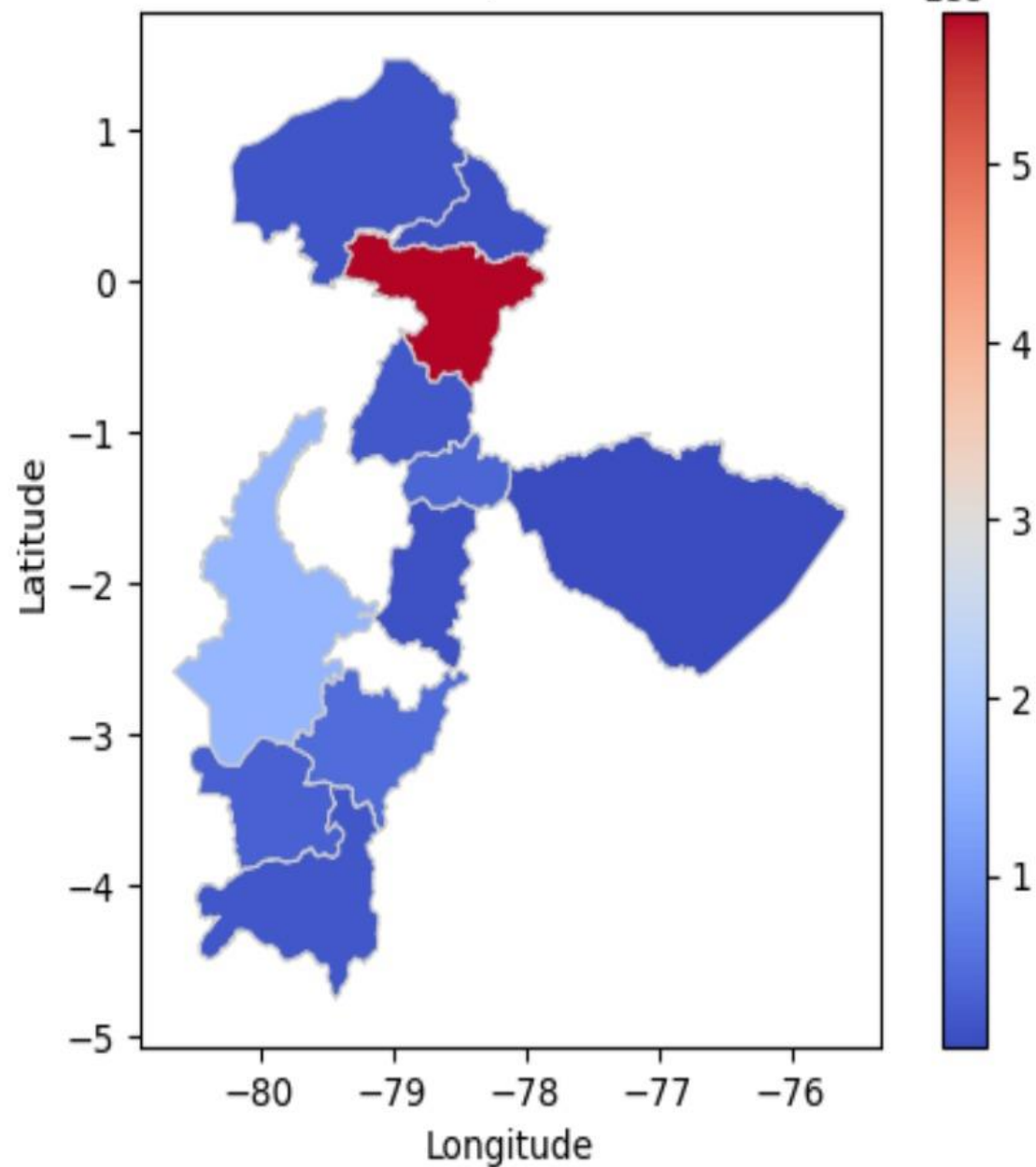
seaborn



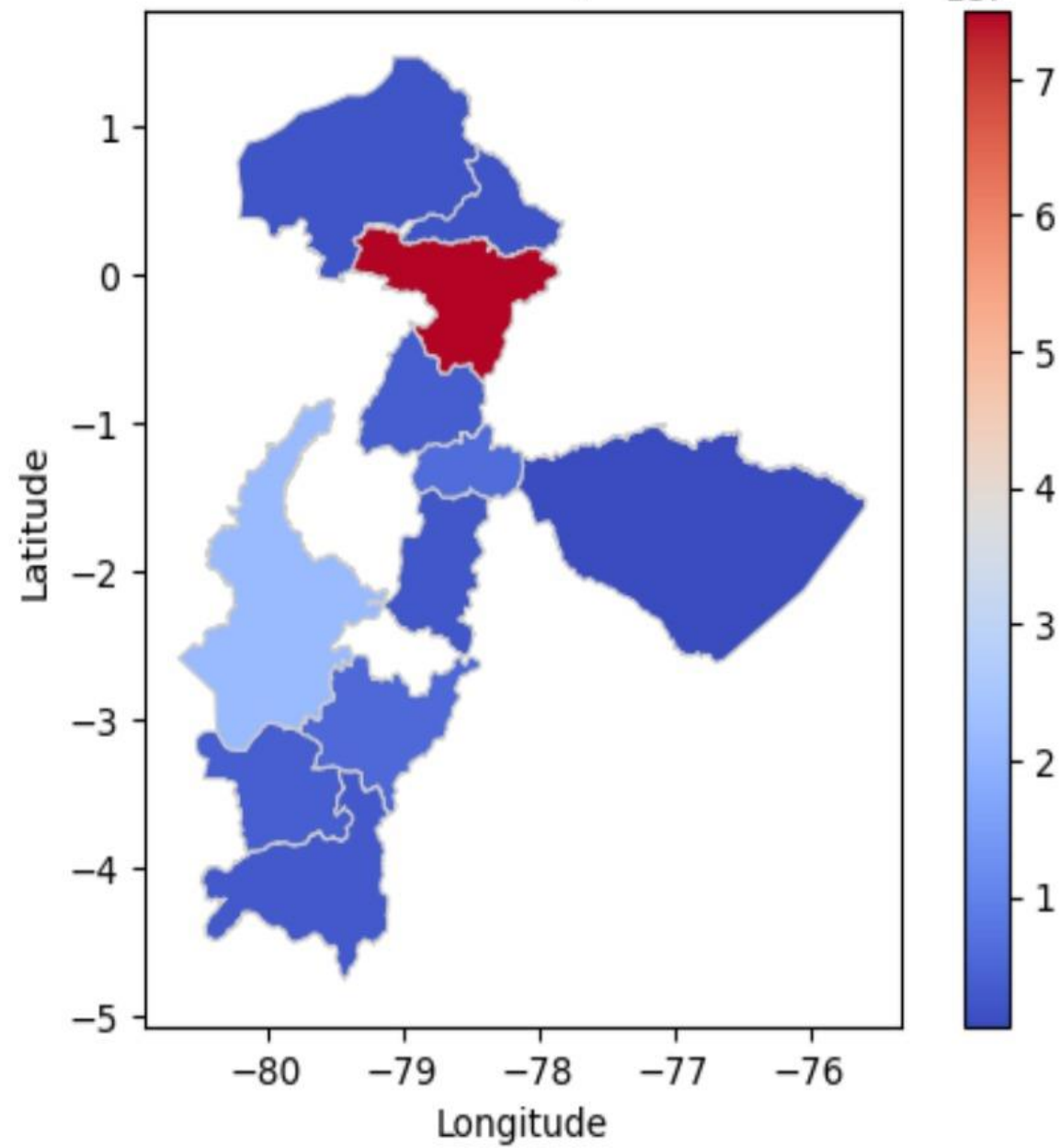
pandas

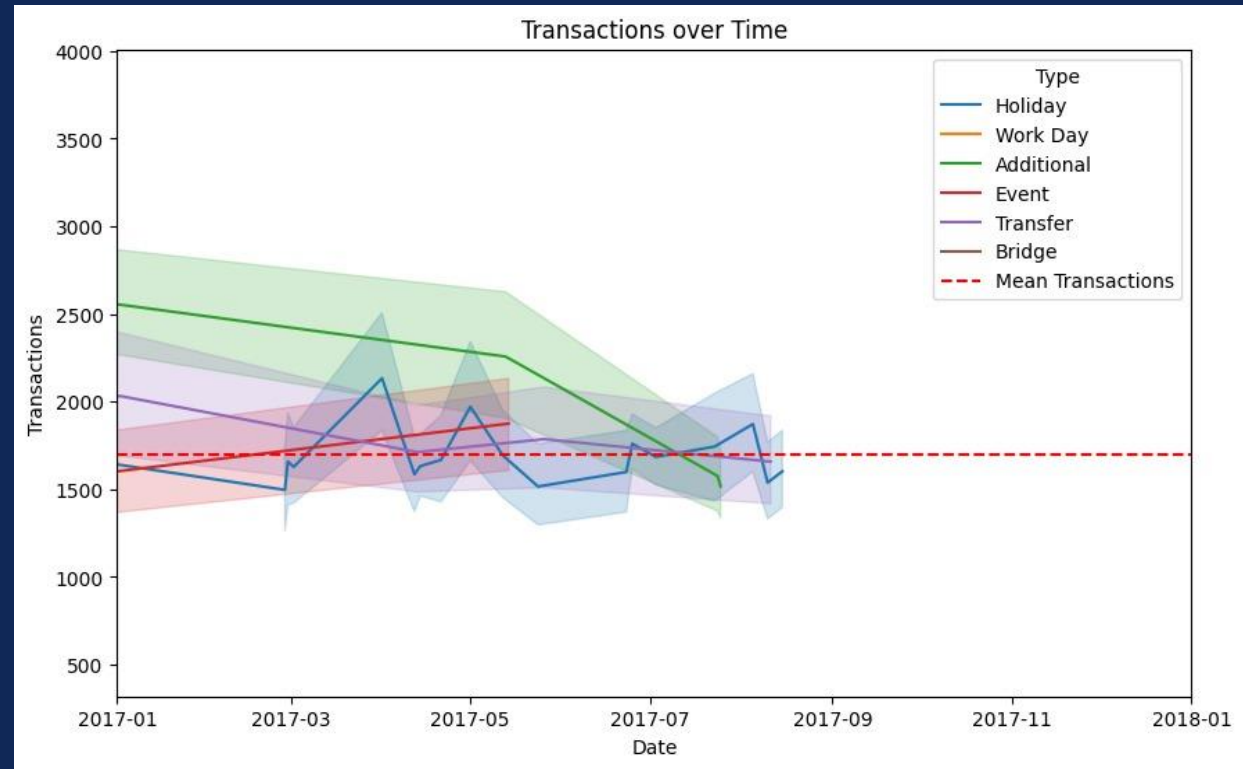
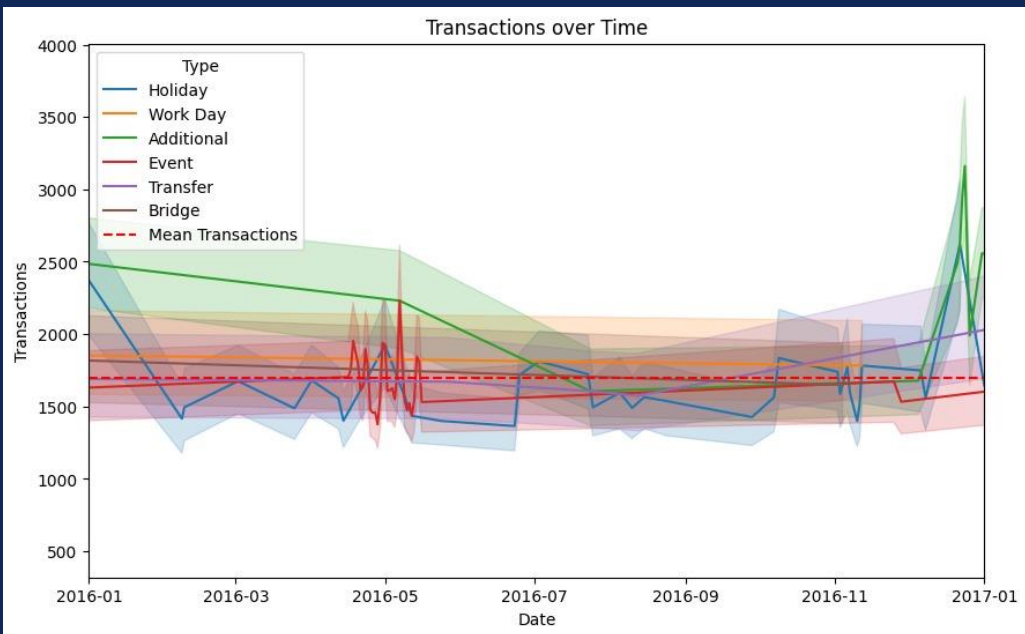
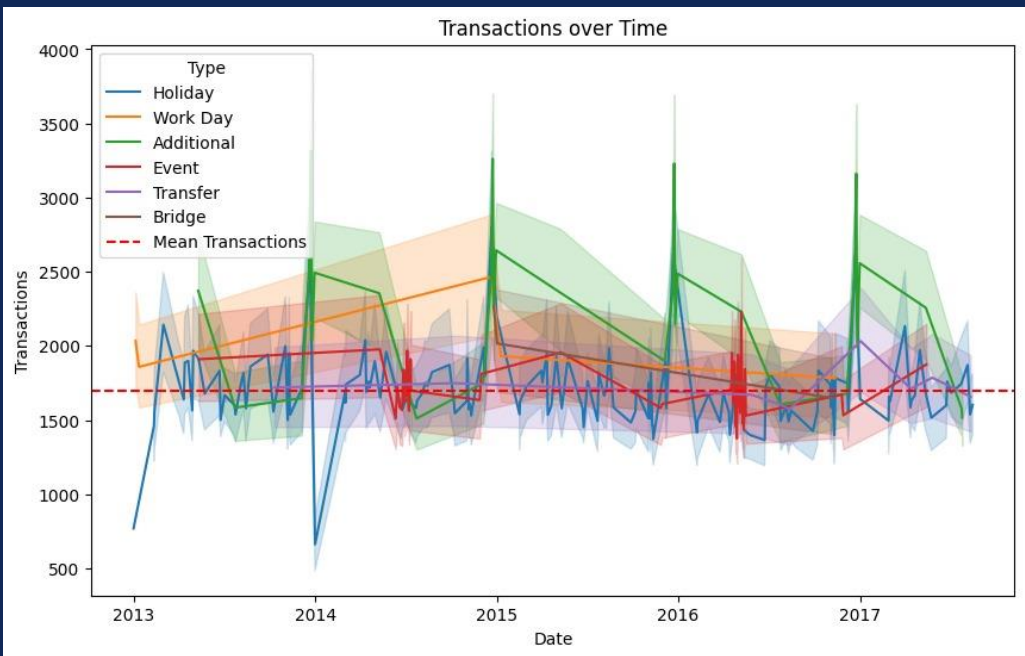


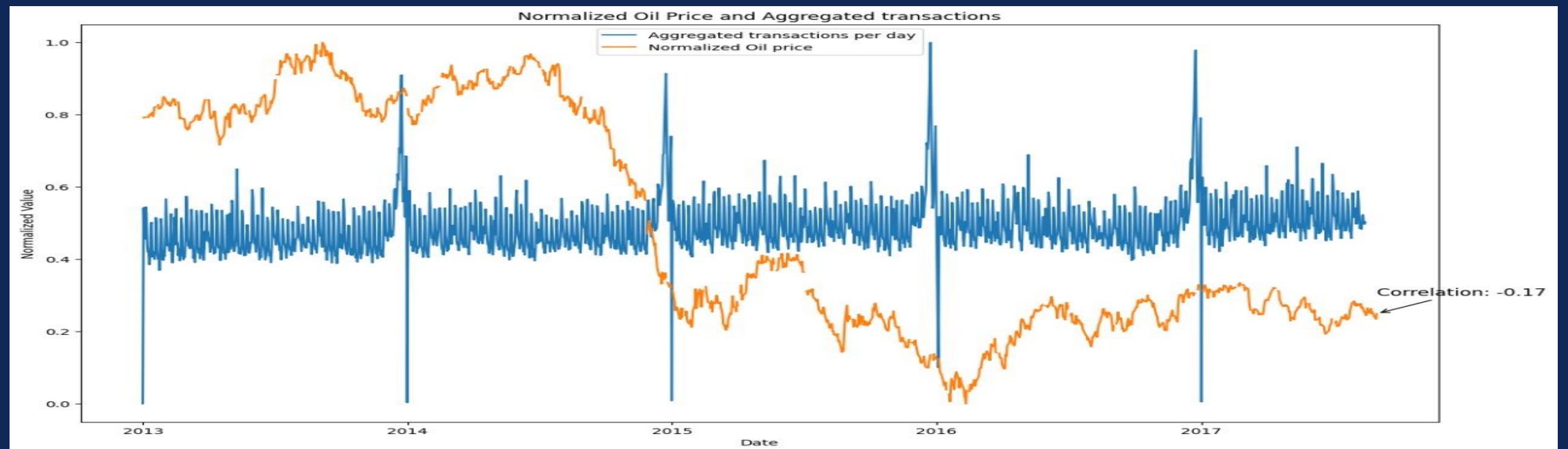
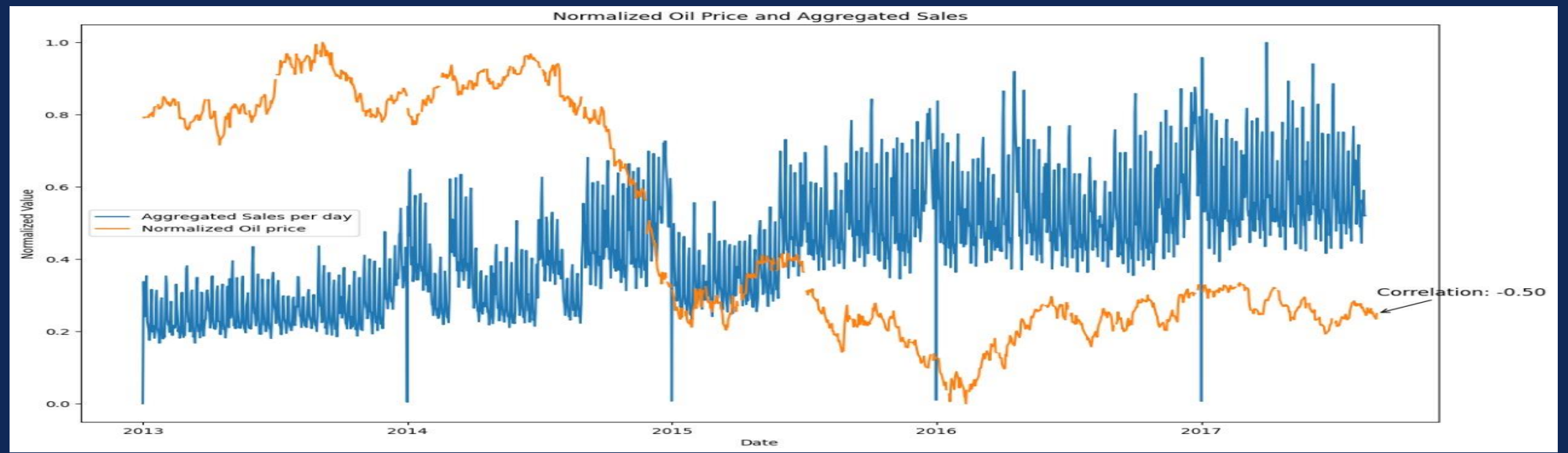
Sales Distribution per State in Ecuador $1e8$



Transactions Distribution per State in Ecuador $1e7$




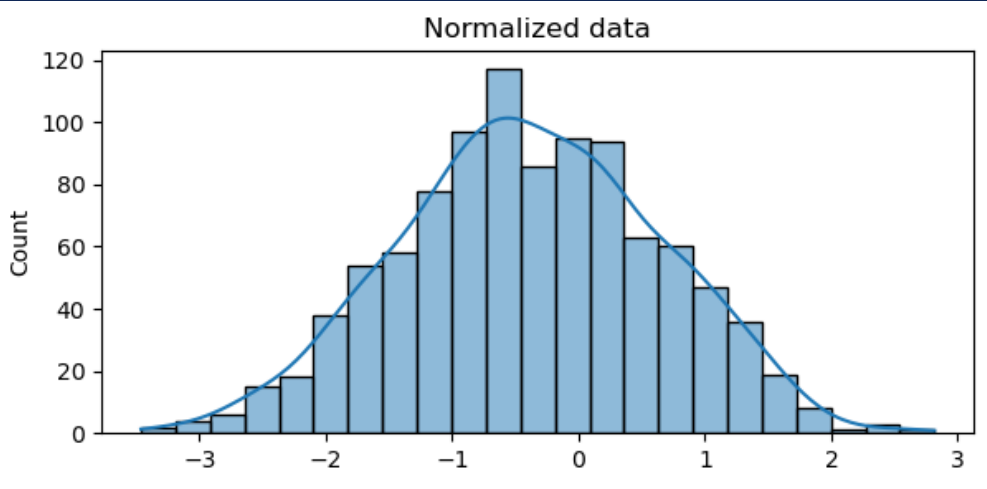
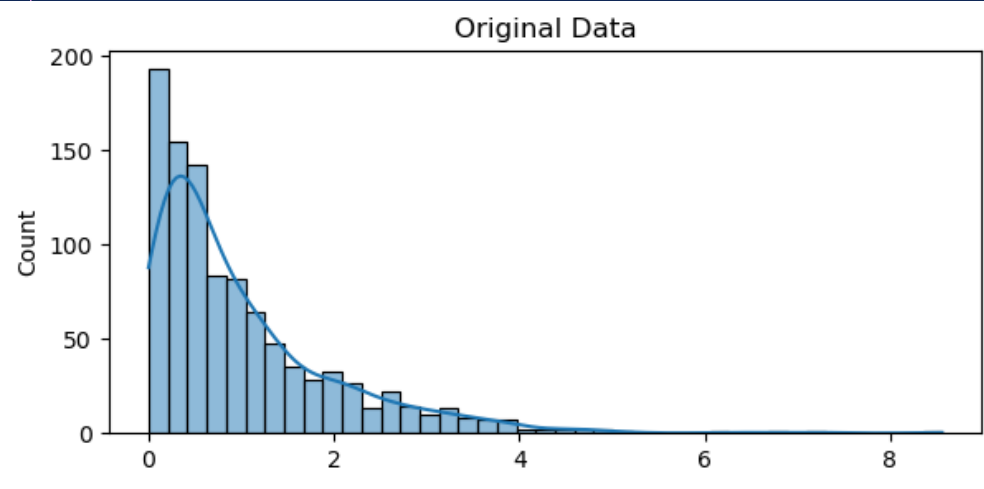




	id	date	store_nbr	family	sales	onpromotion	city	state	store_type	store_cluster	holiday_type	holiday_scope	holiday_g
0	0	2013-01-01	1	AUTOMOTIVE	0.000	0	Quito	Pichincha	D	13	Holiday	National	Eo
1	1	2013-01-01	1	BABY CARE	0.000	0	Quito	Pichincha	D	13	Holiday	National	Eo
2	2	2013-01-01	1	BEAUTY	0.000	0	Quito	Pichincha	D	13	Holiday	National	Eo
3	3	2013-01-01	1	BEVERAGES	0.000	0	Quito	Pichincha	D	13	Holiday	National	Eo
4	4	2013-01-01	1	BOOKS	0.000	0	Quito	Pichincha	D	13	Holiday	National	Eo
...
3000883	3000883	2017-08-15	9	POULTRY	438.133	0	Quito	Pichincha	B	6	Holiday	Local	Riob
3000884	3000884	2017-08-15	9	PREPARED FOODS	154.553	1	Quito	Pichincha	B	6	Holiday	Local	Riob
3000885	3000885	2017-08-15	9	PRODUCE	2419.729	148	Quito	Pichincha	B	6	Holiday	Local	Riob
3000886	3000886	2017-08-15	9	SCHOOL AND OFFICE SUPPLIES	121.000	8	Quito	Pichincha	B	6	Holiday	Local	Riob

ONE HOT
ENCODING
IN PYTHON

color		color_red	color_blue	color_green
red		1	0	0
green		0	0	1
blue		0	1	0
red		1	0	0



ANALISIS MODELOS



Características ▼	Regresion Lineal ▼	Regresion Logistica ▼	Redes Neuronales ▼	Arboles ▼	RandomForests ▼	Clustering ▼
Tipos de Problema	Regresion	Clasificacion	Ambos	Ambos	Ambos	Agrupamiento
Salida	Valor Continuo	Probabilidad	Variable	Clases	Clases	Grupos
Interpretabilidad	Alta	Alta	Baja	Alta	Alta	Moderada
Resistencia Overfitting	Baja	Moderada	Baja(sin regular)	Baja	Alta	Variable
Capacidad Capturar Relaciones NO LINEALES	Baja	Baja	Alta	Alta	Alta	Moderada
Tiempo Entrenamiento	Rápido	Rápido	Varía	Rápido	Moderado	Varía
Requiere Escalado Datos	Sí	Sí	Sí	No	No	A menudo sí
Robustez	No	Moderada	Depende de la arquitectura	No	Sí	Depende del método

DATA

Which dataset do you want to use?



Ratio of training to test data: 50%



Noise: 20



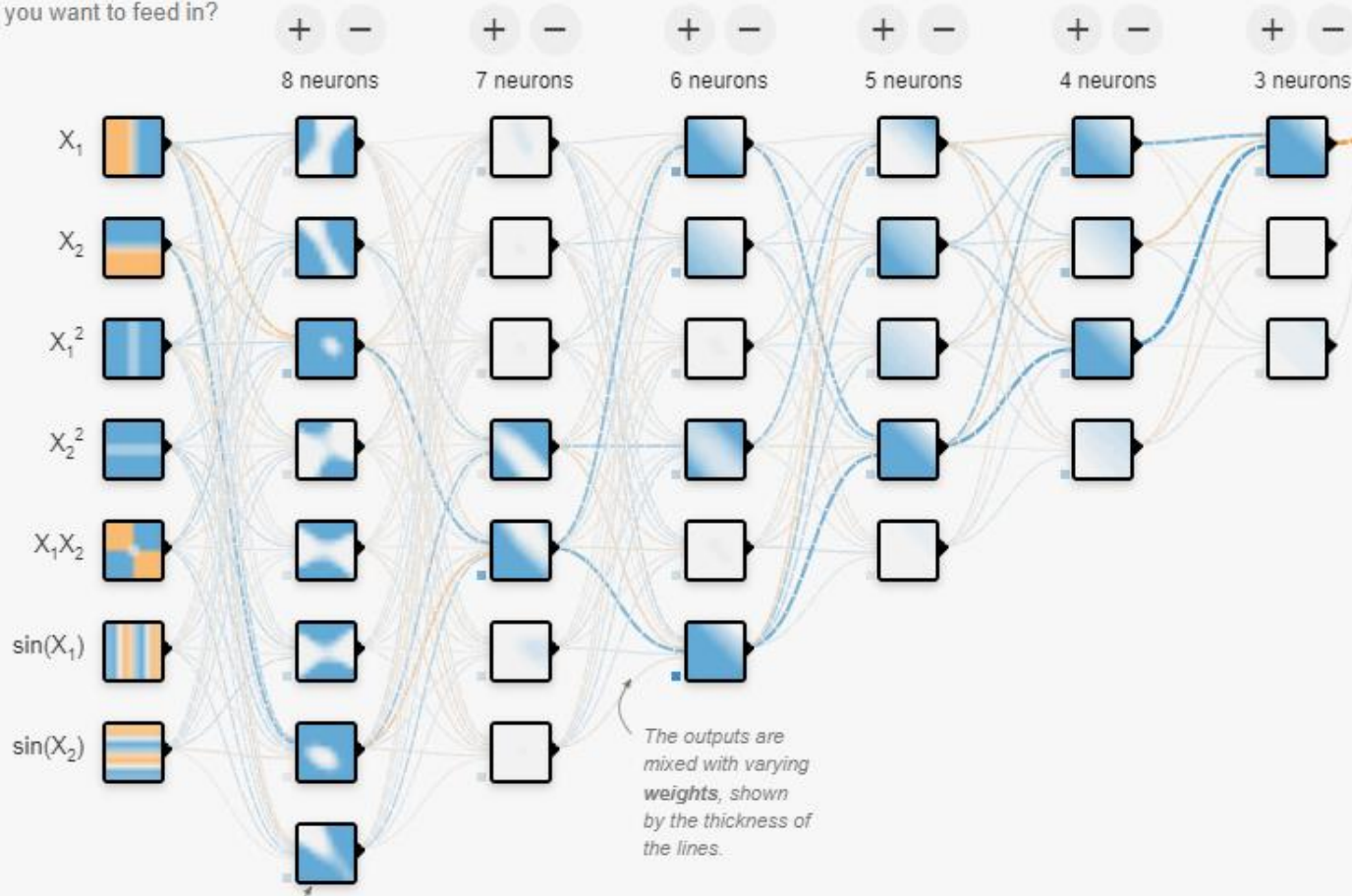
Batch size: 21



REGENERATE

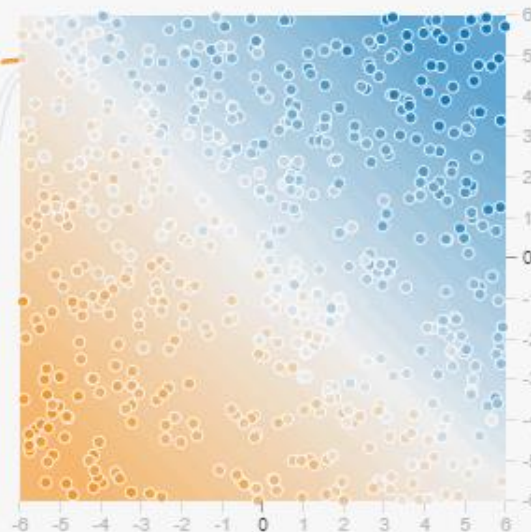
FEATURES

Which properties do you want to feed in?



OUTPUT

Test loss 0.006
Training loss 0.005

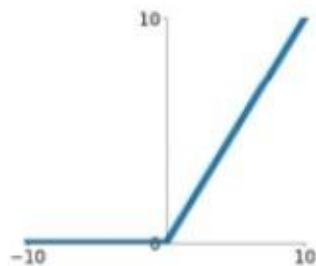


Colors shows data, neuron and weight values.

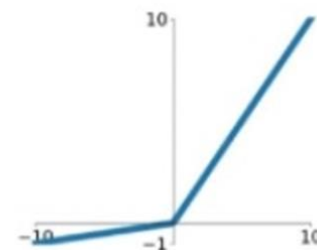


☐ Show test data ☐ Discretize output

ReLU

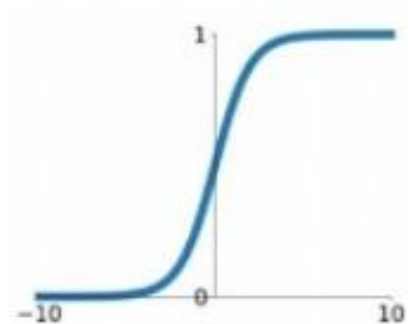
$$\max(0, x)$$


Leaky ReLU

$$\max(0.1x, x)$$


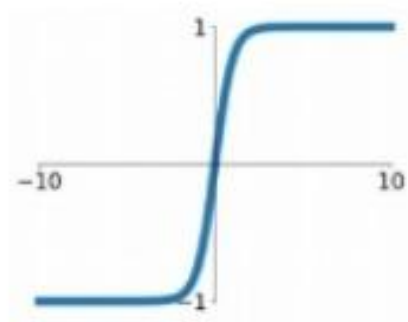
Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



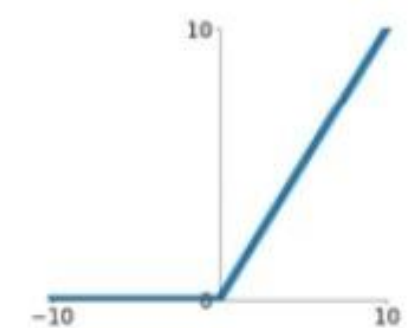
tanh

$$\tanh(x)$$



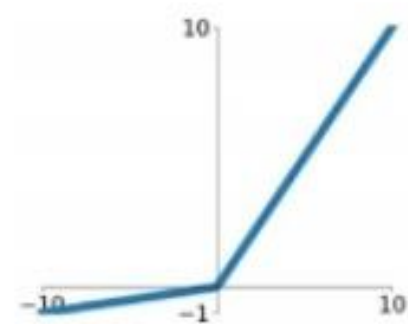
ReLU

$$\max(0, x)$$



Leaky ReLU

$$\max(0.1x, x)$$

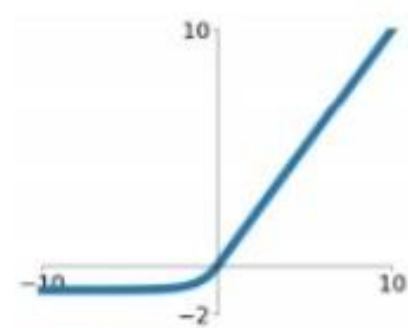


Maxout

$$\max(w_1^T x + b_1, w_2^T x + b_2)$$

ELU

$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$





“METRICAS”

Métrica	Problema de Regresión	Problema de Clasificación	Descripción
Mean Absolute Error (MAE)	Sí	No	Promedio de los errores absolutos entre las predicciones y los verdaderos valores.
Mean Squared Error (MSE)	Sí	No	Promedio de los errores cuadrados entre las predicciones y los verdaderos valores.
Root Mean Squared Error (RMSE)	Sí	No	Raíz cuadrada del MSE.
R ² (Coeficiente de determinación)	Sí	No	Mide cuánta de la variabilidad en el objetivo puede ser explicada por las características.
Accuracy	No	Sí	Proporción de predicciones correctas sobre el total.
Precision	No	Sí	Proporción de verdaderos positivos entre todos los positivos predichos.
Recall (Sensibilidad)	No	Sí	Proporción de verdaderos positivos entre todos los positivos reales.
F1-Score	No	Sí	Media armónica de precisión y recall.
Root Mean Squared Logarithmic Error (RMSLE)	Sí	No	Es útil cuando los errores en las predicciones de valores bajos y altos son igualmente importantes.
Log Loss	No	Sí (especialmente para clasificación binaria)	Penaliza las clasificaciones incorrectas en función de la confianza del modelo.

Es esencial elegir la métrica correcta según el problema .

¿Por qué RMSLE?

- El RMSLE, es menos sensible a los errores grandes en las predicciones en comparación con el RMSE.
- Adaptado para Valores Grandes: Evita penalizaciones excesivas cuando ambos, real y predicho, son números grandes.

PORQUE REESTRUCTURAMOS, EL DATAFRAMES?

Durante las pruebas notamos cambios relevantes al reducir la información por lo tanto empezamos a dirigir el modelo hacia esta nueva idea, cosa que al final logro mejorar los resultados obtenidos

Eficiencia con Grandes Datos: Puede manejar conjuntos de datos extensos rápidamente.

Regularización Integrada: Reduce el sobreajuste, mejorando la generalización.

Optimiza el RMSLE: Entrena el modelo específicamente para la métrica clave de la competencia.

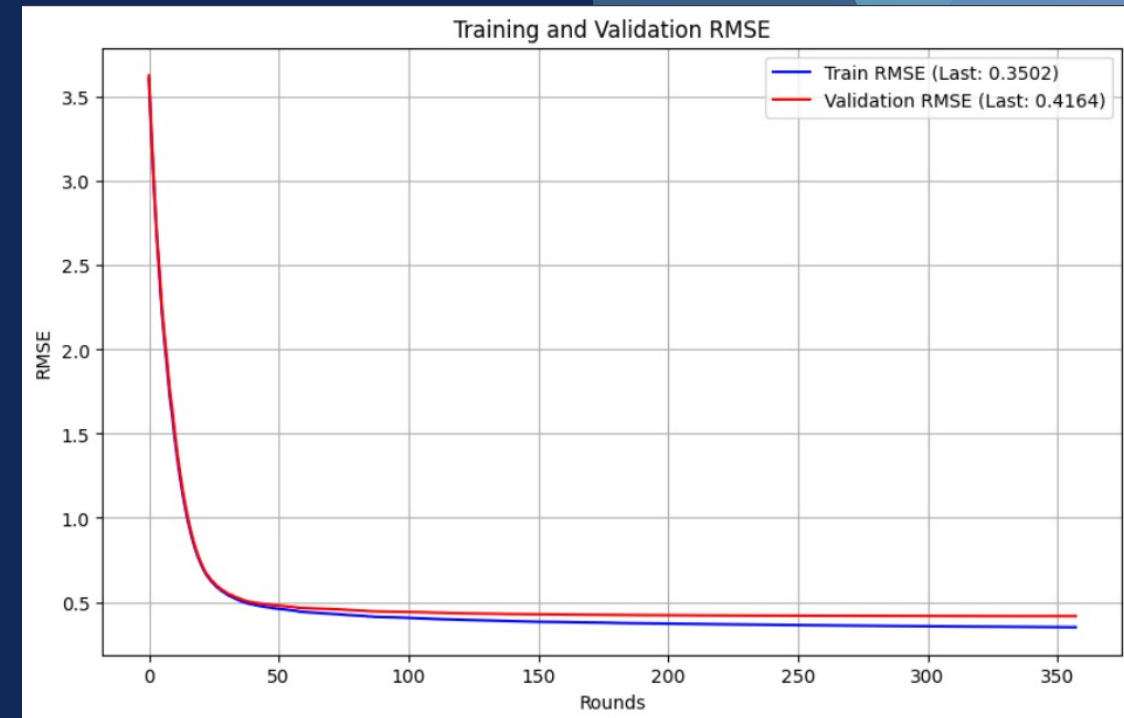
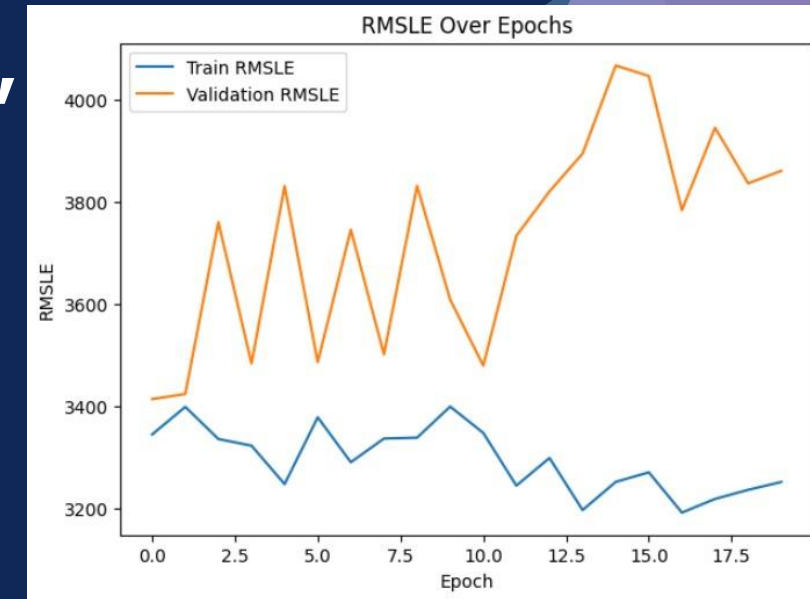
Maneja Valores Faltantes: No requiere imputación previa.

Flexibilidad: Compatible con múltiples bibliotecas y tipos de problemas.

Poda de Árboles: Construye árboles más óptimos evitando excesiva profundidad.

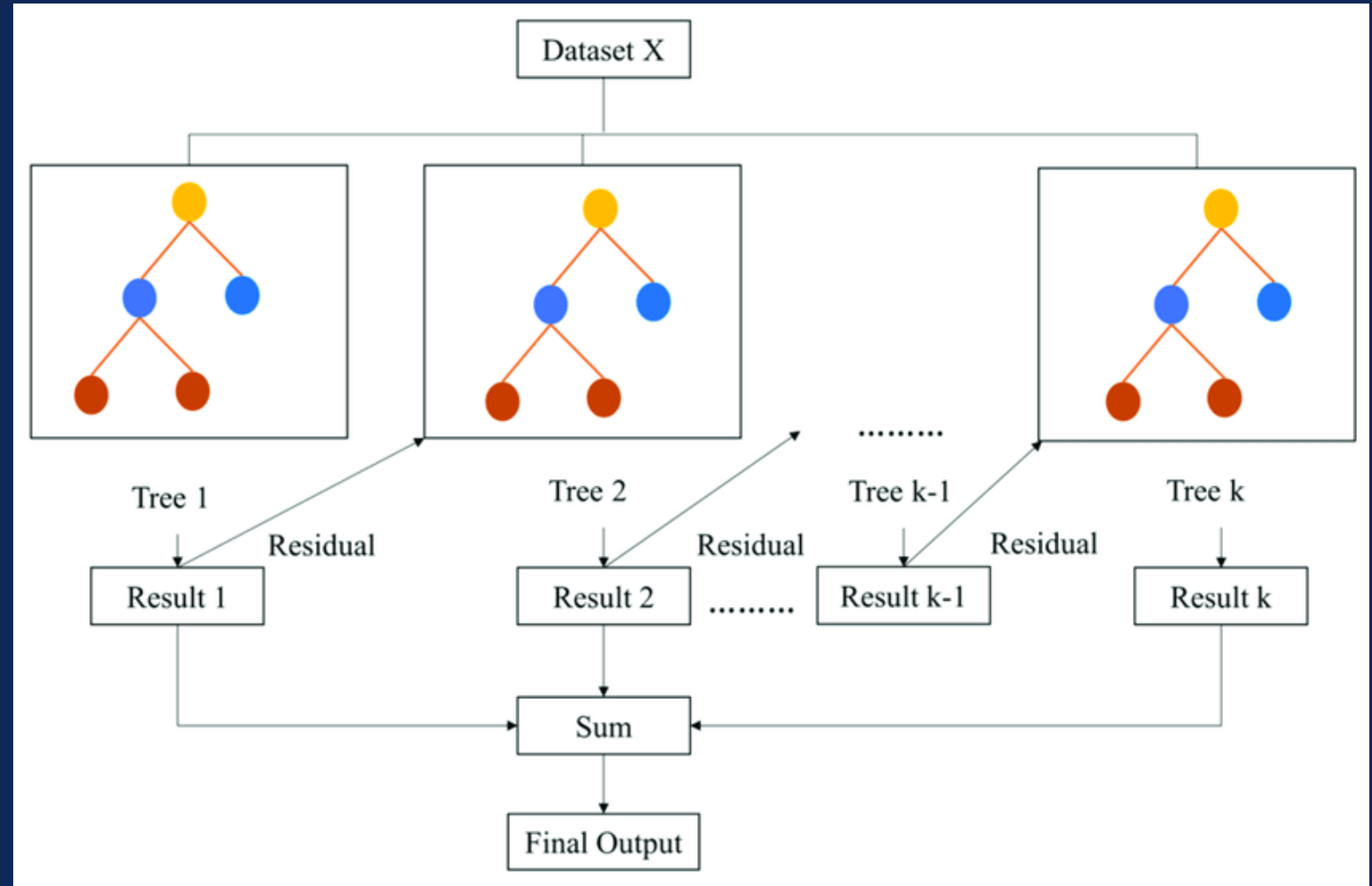
Paralelización: Usa técnicas de paralelización para una construcción de árboles más rápida.

POR ESO...



XGBOOST

Parámetro	Valor
Objective	reg:squarederror
Eval_metric	rmse
Learning rate	0.1
Subsample	0.8
Colsample_bytree	0.8
Reg_alpha	0.01
Reg_lambda	1
Max_depth	10
N_estimators	200
Min_child_weight	47
Random_state	42





RESULTADOS

97

A01275108



0.46554

2

2h



Your Best Entry!

Your most recent submission scored 0.46554, which is an improvement of your previous score of 0.47695. Great job!

[Tweet this](#)

103

Emiliano Mendoza Nieto



0.47046

5

7h



Your Best Entry!

Your most recent submission scored 0.47046, which is an improvement of your previous score of 0.48212. Great job!

[Tweet this](#)Version 1

0.59469

I N T E R F A Z



FLASK API

Sales Prediction

Select Day:	<input type="text" value="1"/>	Select Month:	<input type="text" value="1"/>
Select Type:	<input type="text" value="0"/>		
Select Store Number:	<input type="text" value="0"/>		
Select State:	<input type="text" value="0"/>		
Select Family:	<input type="text" value="0"/>		
Select Onpromotion:	<input type="text" value="0"/>		
Select Weekday:	<input type="text" value="0"/>		
Select Payday:	<input type="text" value="0"/>		
Select City Number:	<input type="text" value="0"/>		
Select Cluster Number:	<input type="text" value="0"/>		
Select Event Name:	<input type="text" value="0"/>		
Select Earthquake:	<input type="text" value="1"/>		
Select Local Holiday Name:	<input type="text" value="0"/>		
Select Regional Holiday Name:	<input type="text" value="0"/>		



THANK YOU
