

CORPORACIÓN FAVORITA

STORE SALES

Renato Meléndez	A01027291
Bruno Salazar	A00825722
Omar Montiel	A00825358
Sebastian Saldaña	A01570274



Análisis del Programa



- 01 Overview del problema
- 02 Hallazgos del análisis estadístico y descriptivo
- 03 Explicación de la metodología usada
- 04 Resultado final
- 05 Siguiendo pasos

Overview del problema

Reto

Construir un modelo de alta precisión que prediga las ventas unitarias de miles de artículos vendidos en diferentes tiendas de Corporación Favorita





Overview del problema

Objetivo

Utilizar time-series forecasting para pronosticar las ventas en tiendas con datos de Corporación Favorita.

Hallazgos del análisis estadístico y descriptivo



 Hallazgos del análisis estadístico y descriptivo

Los datos de Corporación Favorita incluyen fechas, información de tiendas y productos, si el artículo estaba siendo promocionado, así como las cifras de ventas.

Importancia de variables



Días Festivos

El movimiento de días inhábiles por celebración afecta el análisis de días festivos

Terremoto 16 de Abril de 2016

La gente se unió a los esfuerzos de socorro donando agua y otros productos de primera necesidad que afectaron enormemente las ventas de los supermercados durante varias semanas después del terremoto.






Combustible

El comportamiento del precio de combustible en el país ya que Ecuador es dependiente del petróleo y su salud económica es altamente vulnerable a los impactos en los precios del petróleo.

Potencial del proyecto

Los analistas de datos registran las observaciones de datos en intervalos constantes durante un conjunto de períodos de tiempo en lugar de registrar las observaciones de datos al azar. La tasa de observación (intervalo de tiempo) puede ser desde por meses hasta varios años.

-  Corto plazo: mejora en experiencia y fluidez en tiendas
-  Mediano plazo: automatización de promociones a tiempo real
-  Largo plazo: logística avanzada para pedidos grandes



Como se evalúa la solución (RMSLE)



El error logarítmico cuadrático medio (RMSLE) mide la cantidad de error que hay entre dos conjuntos de datos todo esto calculado en una escala logarítmica. En otras palabras, compara un valor predicho y un valor observado o conocido.

 Hallazgos del análisis estadístico y descriptivo

Limpieza de datos

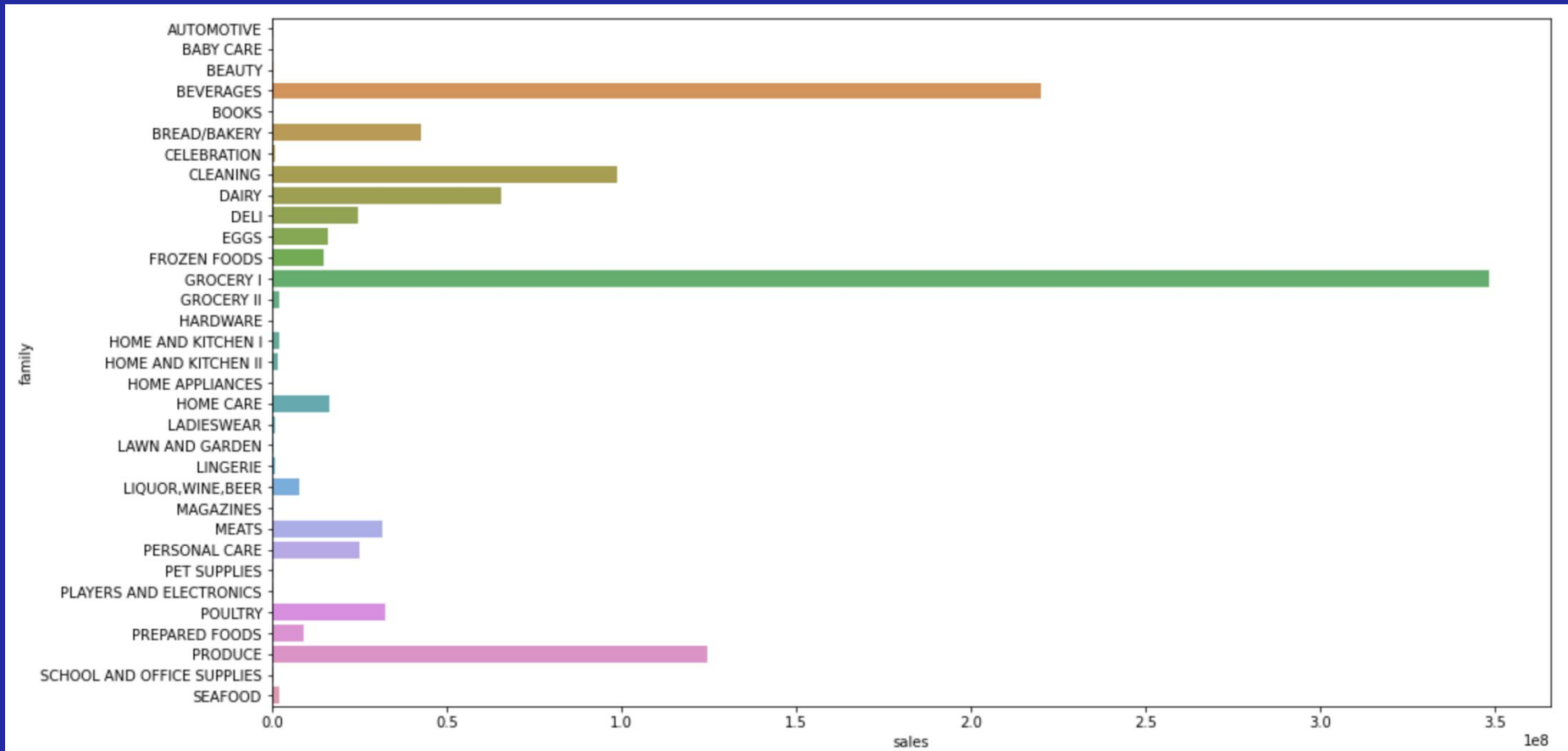
Debido a que la mayoría de las variables tienen los 100% completos, solamente se hizo una ligera limpieza a Días Festivos y a Petróleo. Todas las variables se modificaron para que ahora su índice sea la fecha.



Columna	Descripción
sales	unidades vendidas de la familia de productos
family	familia de producto de la que se hizo la venta

 Hallazgos del análisis estadístico y descriptivo

Ventas por familia



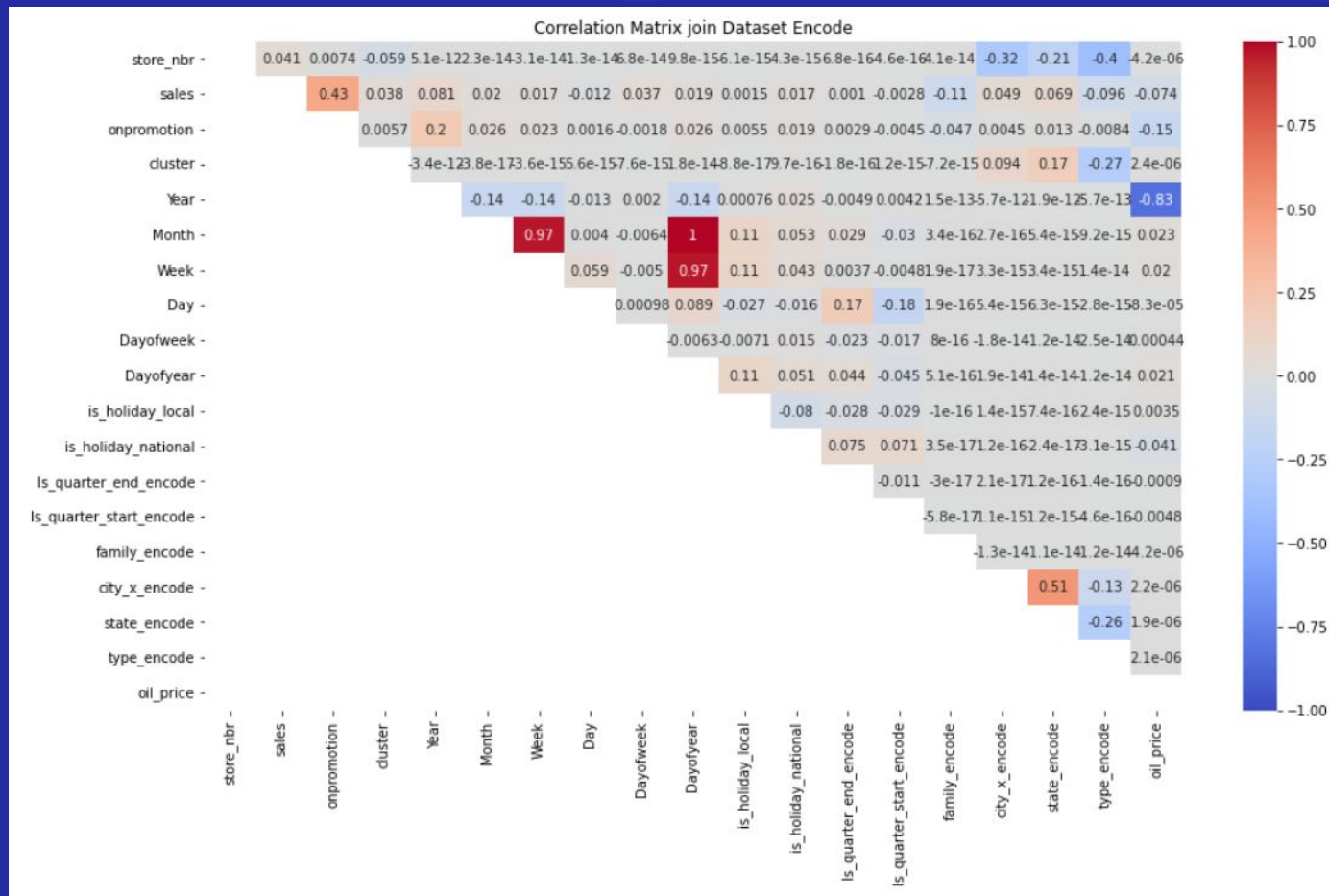
Ventas por familia

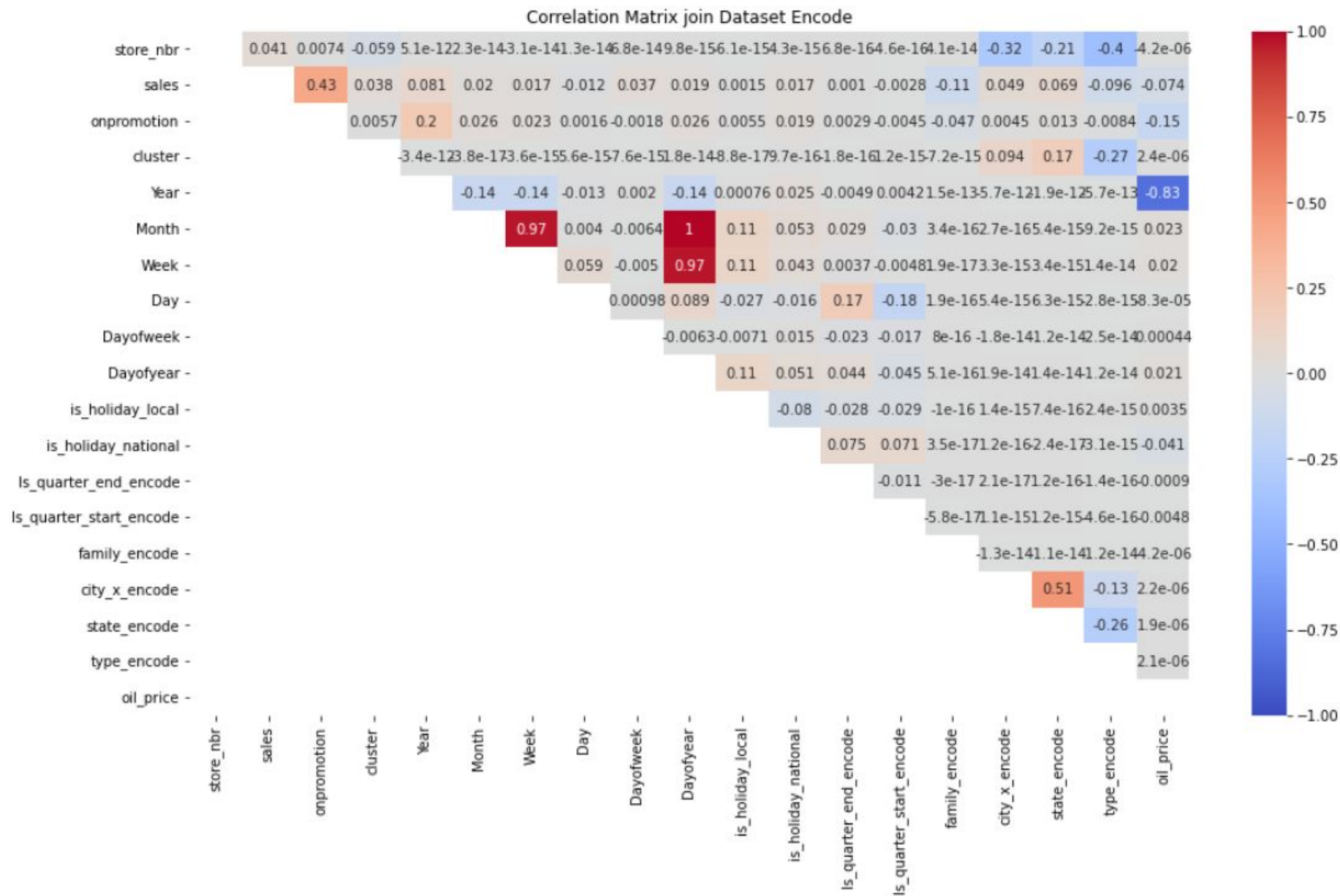


Columna	Descripción
sales	unidades vendidas de la familia de productos
onpromotion	cantidad de artículos de la familia de productos que estaba en promoción
store_nbr	tienda en la que se realizó la venta

 Hallazgos del análisis estadístico y descriptivo

Correlación de Variables





Correlación de Variables



onpromotion
con **sales**



store_nbr
con **sales**

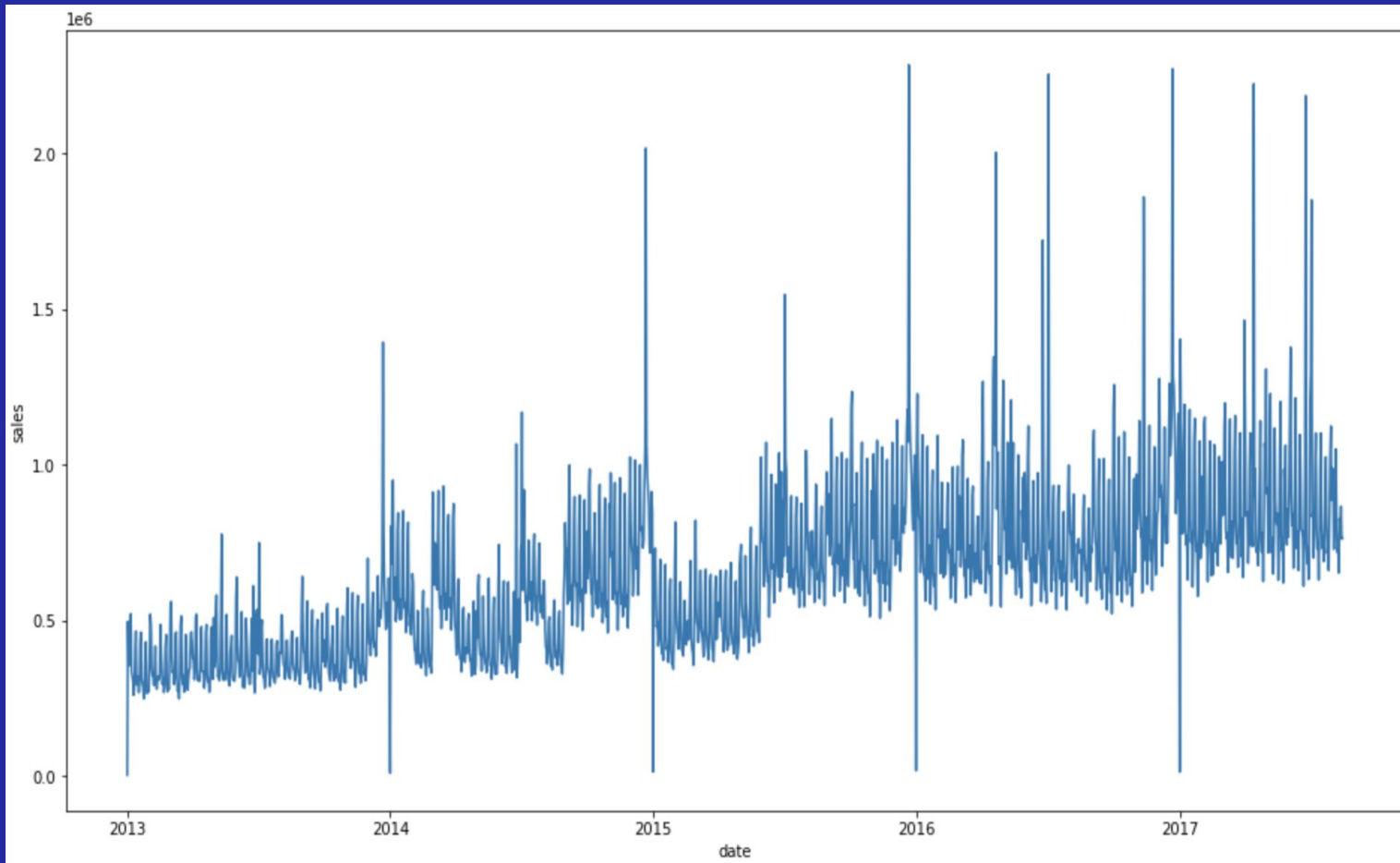


onpromotion con
store_nbr

Columna	Descripción
sales	unidades vendidas de la familia de productos
date	Fecha de cada venta individual

 Hallazgos del análisis estadístico y descriptivo

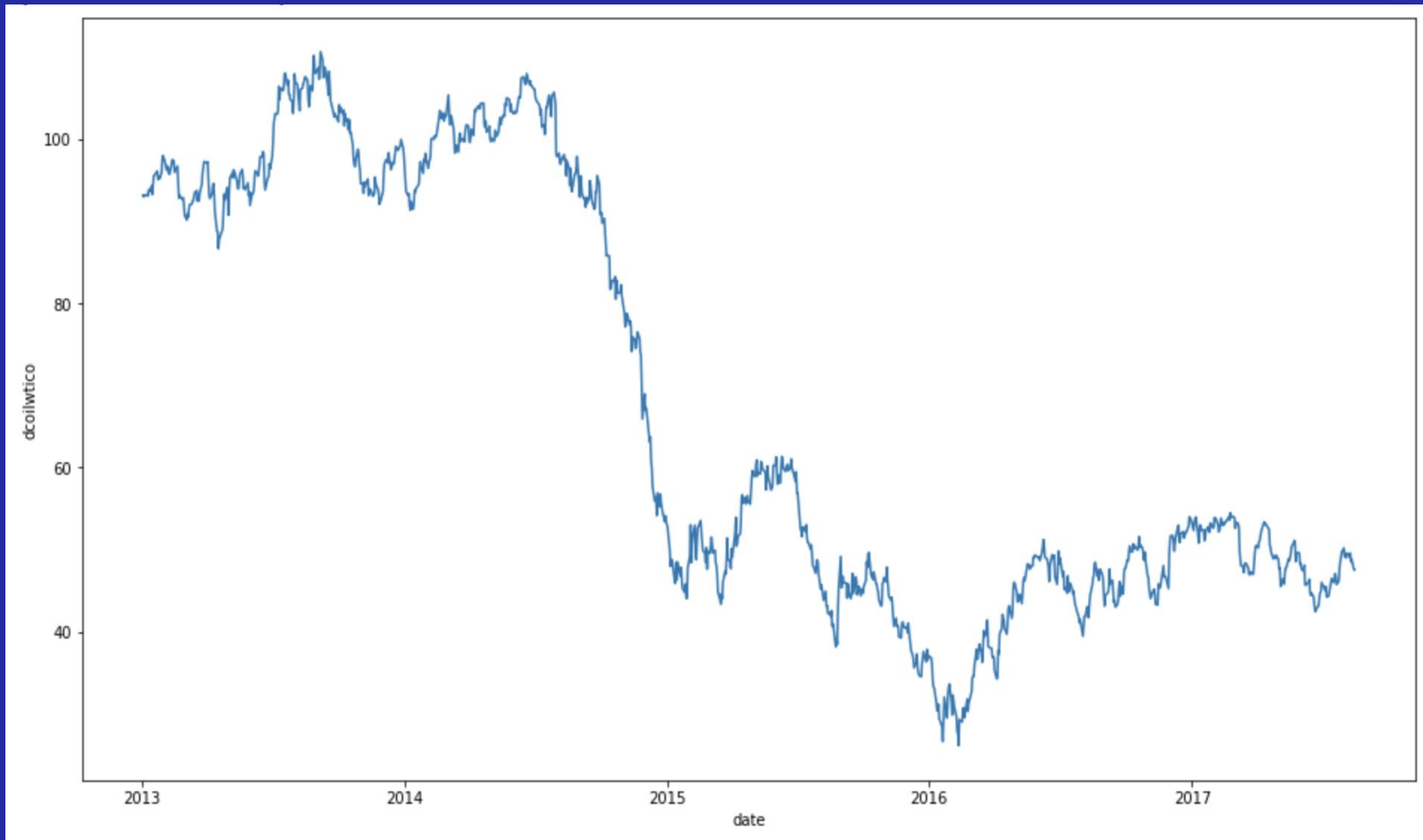
Serie de Tiempo de Ventas



Columna	Descripción
date	Fecha de la cotización del precio
dcoilwtico	precio del petroleo en la fecha

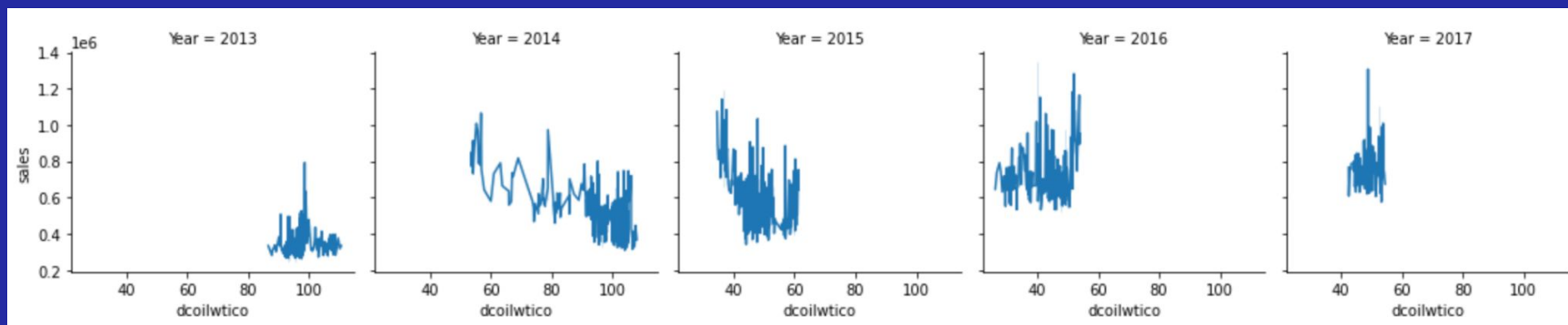
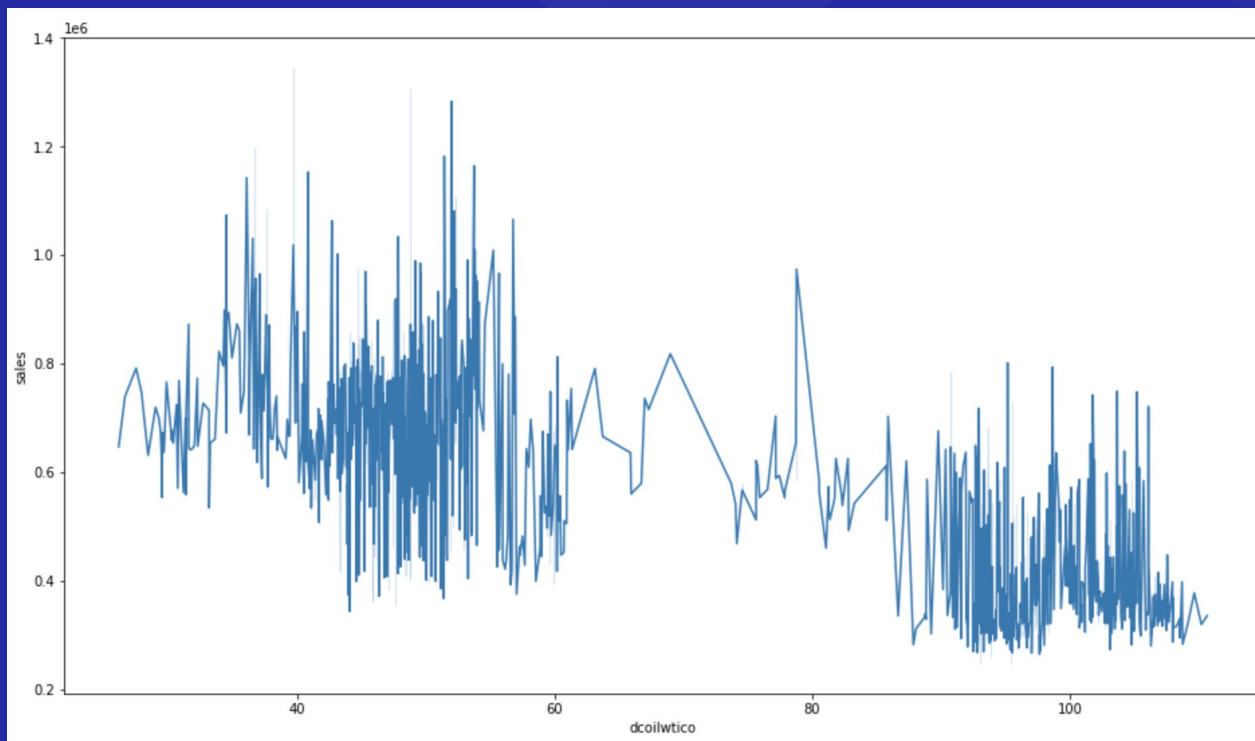
 Hallazgos del análisis estadístico y descriptivo

Precio del petróleo



Ventas y petróleo

Columna	Descripción
date	unidades vendidas de la familia de productos
dcoilwtico	precio del petroleo en la fecha





Explicación de la metodología usada

Feature Engineering

- Unimos la tabla de Stores.
- Descompusimos la fecha en los siguientes campos: Year, Month, Week, Day, Dayofweek, Dayofyear, is_quarter_end, is_quarter_start.
- Posteriormente unimos las holidays, en donde, estas las ajustamos, de acuerdo a si son nacionales o locales.
- Unimos la columna de oil llenando los datos faltantes con un promedio móvil de 7 (semanal).
- Posteriormente aplicamos label encoding en las siguientes variables: 'is_quarter_end', 'is_quarter_start', 'city_x', 'state', 'type'.
- Aplicamos one hot encoding en la columna de Family.



Tareas previas al modelado

- 🍷 Se separó el dataset en un split de 80%-20% para los training y testing sets.
- 🍷 Se definió una función de evaluación que calcula el coeficiente de determinación (R^2) y el error logarítmico de la raíz de la media cuadrada (RMSLE) para los modelos.



Modelaje: Enfoque 1

El primero realizando todo lo descrito en feature engineering menos one hot encoding, en donde se utilizaron 10 algoritmos:

Linear Regression, Ada Boost Regression, Bagging Regression, Gradient Boosting Regression, Extra Trees Regression, Histogram Gradient Boosting Regression, Linear Support Vector Machine Regression, Decision Trees Regression, Multi Layer Perceptron Regression, Random Forest Regression.

Los 4 mejores fueron (RMSLE):

 Bagging Regression	0.4671
 Extra Tree Regression	0.5056
 Decision Tree Regression	0.5354
 Random Forest Regression	0.4665

Enfoque adicional

Después de ver los resultados del *testing* del primer enfoque, se intentó utilizar el mismo pero solamente con las features que tuvieran una mayor correlación absoluta respecto a la variable de *sales*. Esto se ejecutó con los cuatro regresores que tuvieron menor RMSLE en esta primera iteración.

Al eliminar la mayoría de las variables, los resultados no mostraron una mejora (pasaron de ~0.5 a ~0.9 en el *training set*):

```
Results for BR on normalmode
Training score BR-normalmode with score: -86226.76042189564
Test scores
explained_variance: 0.7807
r2: 0.7807
root_mean_squared_log_error: 0.9309
```

```
Results for ETR on normalmode
Training score ETR-normalmode with score: -66523.42831261999
Test scores
explained_variance: 0.7665
r2: 0.7665
root_mean_squared_log_error: 0.9938
```

```
Results for DT on normalmode
Training score DT-normalmode with score: -66523.40102515776
Test scores
explained_variance: 0.7492
r2: 0.7491
root_mean_squared_log_error: 0.9954
```

```
Results for RF on normalmode
Training score RF-normalmode with score: -86578.53767473513
Test scores
explained_variance: 0.7809
r2: 0.7809
root_mean_squared_log_error: 0.9324
```

```
['Normal_y_pred.joblib']
```

Modelaje: Enfoque 2

El segundo utilizando la librería de skforecast

En donde se realizó un forecast autorregresivo recursivo utilizando el algoritmo de Random Forest Regressor.

Donde hicimos aproximadamente 1,782 dataframes puesto que calculamos por tienda (54) y por familia (33).

Utilizando la variables exógenas de:

- oil_price.
- is_holiday_national.
- is_holiday_local.

Una variable exógena es aquella que sabemos su valor en el futuro de manera anticipada.

Modelaje: Enfoque 3 (La última y nos vamos)

Aquí realizamos el mismo enfoque que el primero solo que ahora utilizando one hot encoding en la columna de familias utilizando.

Posteriormente utilizamos los algoritmos:

Linear Regression 2.5161

Ada Boost Regression 3.8365

Random Forest Regression 0.4513

Optando por utilizar el Random Forest Regression



Resultado **Final**

Modelaje: Enfoque 1



633

Bernardo Salazar



3.57565

1

1s



Your First Entry!

Welcome to the leaderboard!

Modelaje: Enfoque 2

620

Bernardo Salazar



2.40494

2

1h

Modelaje: Enfoque 3



YOUR RECENT SUBMISSION



espero_que_salga_bien.csv

Submitted by mrbern · Submitted just now

Score: 0.54731

↓ [Jump to your leaderboard position](#)

Siguientes pasos



El futuro de Corporación Favorita

🔥 Ajuste de hiperparámetros:

Para el Forecast Autorregresivo Recursivo se utilizó un lag de 0.8 y de 0.2, pero no probamos que estos hayan sido los mejores valores. Con un Grid Search Forecaster podríamos ajustar estos valores y potencialmente reducir el error de predicción.

🔥 Utilizar Forecast Directo Multi-Step:

Esta metodología de Forecast no fue utilizada, pero hay datasets que muestran tener un mejor rendimiento con este modelo, por lo que podría ser útil intentarlo.

🔥 Utilizar Grid Search CV:

En el modelo ganador con cross-validation.

Gracias por su
Atención

