Loading datasets...

Train data shape: (3000888, 6)

Test data shape: (28512, 5)

Sample submission shape: (28512, 2)

Missing values in train: 0

Missing values in test: 0

Missing values in oil: date 0

dcoilwtico 43

dtype: int64

Handling missing oil prices...

After filling, missing values in oil: date 0

dcoilwtico 0

dtype: int64

Exploring the data...

Training data date range: 2013-01-01 00:00:00 to 2017-08-15 00:00:00

Test data date range: 2017-08-16 00:00:00 to 2017-08-31 00:00:00

Number of days in test data: 16

Store types distribution:

type

D 18

C 15

A 9

B 8

E 4

Name: count, dtype: int64

Product families:

33

family

AUTOMOTIVE 90936

HOME APPLIANCES 90936

SCHOOL AND OFFICE SUPPLIES 90936

PRODUCE 90936

PREPARED FOODS 90936

Name: count, dtype: int64

Creating visualizations...

Jan 2013Jul 2013Jan 2014Jul 2014Jan 2015Jul 2015Jan 2016Jul 2016Jan 2017Jul 201700.2M0.4M0.6M0.8M1M1.2M1.4M

Total Sales Over TimeDateTotal Sales

Jan 2013Jul 2013Jan 2014Jul 2014Jan 2015Jul 2015Jan 2016Jul 2016Jan 2017Jul 20170100k200k300k400k500k

typeABCDESales by Store TypeDateTotal Sales

Jan 2013Jul 2013Jan 2014Jul 2014Jan 2015Jul 2015Jan 2016Jul 2016Jan 2017Jul 201700.2M0.4M0.6M0.8M1M1.2M1.4M30405060708090100110

SalesOil PriceSales and Oil Price Over TimeDateSalesOil Price (USD)

HolidayRegular Day00.2M0.4M0.6M0.8M1M1.2M1.4M

Sales: Holidays vs Regular DaysDay TypeTotal Sales

Preprocessing data for modeling...

Prepared daily sales dataset:

sales dcoilwtico is\_holiday dayofweek month year

date

2013-01-01 2511.618999 93.14 1 1 1 2013

2013-01-02 496092.417944 93.14 0 2 1 2013

2013-01-03 361461.231124 92.97 0 3 1 2013

2013-01-04 354459.677093 93.12 0 4 1 2013

2013-01-05 477350.121229 93.12 1 5 1 2013

Performing time series analysis...

ADF Statistic: -2.6161957486048584

p-value: 0.08969592175787444

Critical Values: {'1%': -3.4342930672797065, '5%': -2.8632816397229064, '10%': -2.567697207194407}

The series is not stationary

After differencing:

ADF Statistic: -11.49579840542527

p-value: 4.617675539031655e-21

The differenced series is stationary

00.5M1M1.5M0.4M0.6M0.8M1M1.2M−50k050kJan 2013Jul 2013Jan 2014Jul 2014Jan 2015Jul 2015Jan 2016Jul 2016Jan 2017Jul 2017−1M−0.5M00.5M

ObservedTrendSeasonalResidualTime Series Decomposition

Forecast dates:

DatetimeIndex(['2017-08-16', '2017-08-17', '2017-08-18', '2017-08-19',

'2017-08-20', '2017-08-21', '2017-08-22', '2017-08-23',

'2017-08-24', '2017-08-25', '2017-08-26', '2017-08-27',

'2017-08-28', '2017-08-29', '2017-08-30'],

dtype='datetime64[ns]', freq='D')

Prepared forecast dataframe:

dayofweek month year dcoilwtico is\_holiday

2017-08-16 2 8 2017 46.80 0

2017-08-17 3 8 2017 47.07 0

2017-08-18 4 8 2017 48.59 0

2017-08-19 5 8 2017 48.59 0

2017-08-20 6 8 2017 48.59 0

Training ARIMA model...

SARIMAX Results

==============================================================================

Dep. Variable: sales No. Observations: 1684

Model: ARIMA(1, 1, 1) Log Likelihood -22336.996

Date: Sat, 15 Mar 2025 AIC 44679.992

Time: 11:24:11 BIC 44696.277

Sample: 0 HQIC 44686.024

- 1684

Covariance Type: opg

==============================================================================

coef std err z P>|z| [0.025 0.975]

------------------------------------------------------------------------------

ar.L1 0.4419 0.018 24.285 0.000 0.406 0.478

ma.L1 -0.9609 0.008 -124.779 0.000 -0.976 -0.946

sigma2 2.467e+10 1.53e-13 1.62e+23 0.000 2.47e+10 2.47e+10

===================================================================================

Ljung-Box (L1) (Q): 29.35 Jarque-Bera (JB): 1325.63

Prob(Q): 0.00 Prob(JB): 0.00

Heteroskedasticity (H): 2.58 Skew: 0.21

Prob(H) (two-sided): 0.00 Kurtosis: 7.33

===================================================================================

Warnings:

[1] Covariance matrix calculated using the outer product of gradients (complex-step).

[2] Covariance matrix is singular or near-singular, with condition number inf. Standard errors may be unstable.

ARIMA forecast:

1684 804625.851702

1685 823169.368373

1686 831363.599161

1687 834984.563803

1688 836584.638938

1689 837291.699264

1690 837604.143532

1691 837742.210140

1692 837803.220662

1693 837830.180722

1694 837842.094155

1695 837847.358605

1696 837849.684923

1697 837850.712904

1698 837851.167161

Name: predicted\_mean, dtype: float64

Training SARIMA model...

NaN values in dcoilwtico: 0

NaN values in is\_holiday: 0

DEBUG:cmdstanpy:input tempfile: /tmp/tmp7ym3mrn\_/7xevx5x6.json

DEBUG:cmdstanpy:input tempfile: /tmp/tmp7ym3mrn\_/28c71d91.json

DEBUG:cmdstanpy:idx 0

DEBUG:cmdstanpy:running CmdStan, num\_threads: None

DEBUG:cmdstanpy:CmdStan args: ['/usr/local/lib/python3.11/dist-packages/prophet/stan\_model/prophet\_model.bin', 'random', 'seed=69409', 'data', 'file=/tmp/tmp7ym3mrn\_/7xevx5x6.json', 'init=/tmp/tmp7ym3mrn\_/28c71d91.json', 'output', 'file=/tmp/tmp7ym3mrn\_/prophet\_modelqmyzzt0m/prophet\_model-20250315112415.csv', 'method=optimize', 'algorithm=lbfgs', 'iter=10000']

11:24:15 - cmdstanpy - INFO - Chain [1] start processing

INFO:cmdstanpy:Chain [1] start processing

SARIMAX Results

=========================================================================================

Dep. Variable: sales No. Observations: 1684

Model: SARIMAX(1, 1, 1)x(1, 1, 1, 7) Log Likelihood -21781.421

Date: Sat, 15 Mar 2025 AIC 43576.842

Time: 11:24:15 BIC 43614.811

Sample: 0 HQIC 43590.908

- 1684

Covariance Type: opg

==============================================================================

coef std err z P>|z| [0.025 0.975]

------------------------------------------------------------------------------

dcoilwtico 1658.6298 2062.835 0.804 0.421 -2384.452 5701.712

is\_holiday -2046.9887 1.09e+04 -0.187 0.851 -2.35e+04 1.94e+04

ar.L1 0.4110 0.023 17.681 0.000 0.365 0.457

ma.L1 -0.9112 0.019 -46.795 0.000 -0.949 -0.873

ar.S.L7 -0.0227 0.034 -0.669 0.503 -0.089 0.044

ma.S.L7 -0.7783 0.029 -26.590 0.000 -0.836 -0.721

sigma2 1.765e+10 0.077 2.28e+11 0.000 1.77e+10 1.77e+10

===================================================================================

Ljung-Box (L1) (Q): 0.00 Jarque-Bera (JB): 43121.96

Prob(Q): 0.94 Prob(JB): 0.00

Heteroskedasticity (H): 2.80 Skew: -1.21

Prob(H) (two-sided): 0.00 Kurtosis: 27.73

===================================================================================

Warnings:

[1] Covariance matrix calculated using the outer product of gradients (complex-step).

[2] Covariance matrix is singular or near-singular, with condition number 1.97e+26. Standard errors may be unstable.

SARIMA forecast:

1684 7.665016e+05

1685 6.336693e+05

1686 7.753480e+05

1687 9.246088e+05

1688 1.012018e+06

1689 7.728705e+05

1690 7.354000e+05

1691 7.512644e+05

1692 6.197621e+05

1693 7.642938e+05

1694 9.101701e+05

1695 9.975709e+05

1696 7.615290e+05

1697 7.247293e+05

1698 7.381988e+05

Name: predicted\_mean, dtype: float64

Training Prophet model...

NaN values in Prophet data before cleaning:

ds 0

y 0

oil\_price 0

holiday 0

dtype: int64

NaN values in Prophet data after cleaning:

ds 0

y 0

oil\_price 0

holiday 0

dtype: int64

Fitting Prophet model with regressors...

11:24:15 - cmdstanpy - INFO - Chain [1] done processing

INFO:cmdstanpy:Chain [1] done processing

Creating future dataframe...

Adding regressors to future dataframe...

Making Prophet forecast...

Prophet forecast:

ds yhat yhat\_lower yhat\_upper

1684 2017-08-16 7.819043e+05 645451.195052 9.161269e+05

1685 2017-08-17 6.977761e+05 563111.169019 8.252923e+05

1686 2017-08-18 7.783083e+05 649198.344263 9.100587e+05

1687 2017-08-19 9.734944e+05 843653.374490 1.104586e+06

1688 2017-08-20 1.032491e+06 894541.324454 1.177151e+06

1689 2017-08-21 8.251408e+05 681881.764158 9.704445e+05

1690 2017-08-22 7.842395e+05 646650.339477 9.175259e+05

1691 2017-08-23 8.153090e+05 686585.530158 9.482534e+05

1692 2017-08-24 7.539858e+05 613455.022487 8.939797e+05

1693 2017-08-25 8.106676e+05 672165.511636 9.453461e+05

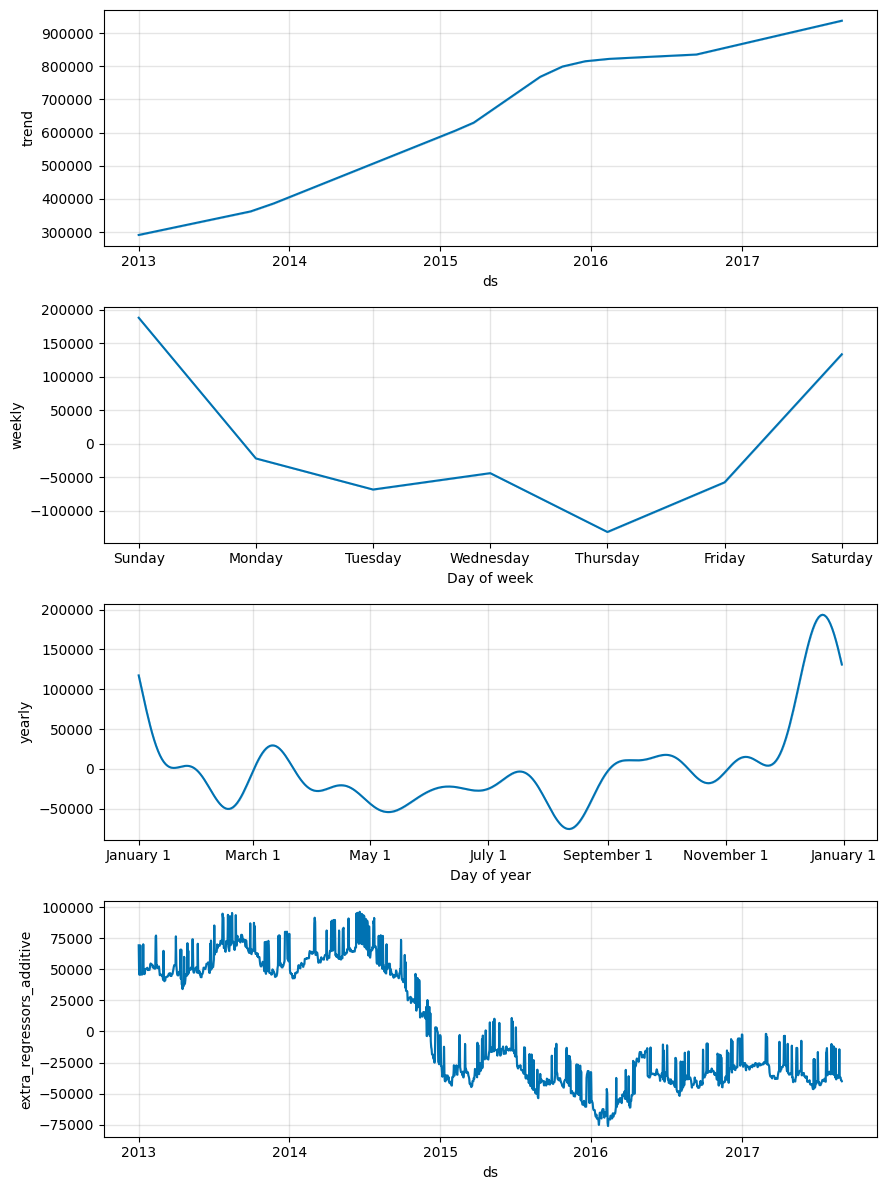
1694 2017-08-26 1.006924e+06 867451.685977 1.148096e+06

1695 2017-08-27 1.066470e+06 931884.933432 1.203233e+06

1696 2017-08-28 8.590683e+05 727379.515898 9.888667e+05

1697 2017-08-29 8.173502e+05 684584.926249 9.566260e+05

1698 2017-08-30 8.451380e+05 707923.609023 9.751944e+05



Jan 2013Jul 2013Jan 2014Jul 2014Jan 2015Jul 2015Jan 2016Jul 2016Jan 2017Jul 201700.2M0.4M0.6M0.8M1M1.2M1.4M

ModelsHistorical SalesARIMA ForecastSARIMA ForecastProphet ForecastProphet Confidence IntervalSales Forecasts Comparison: ARIMA, SARIMA, and ProphetDateSales

Preparing submission file...

Creating ensemble forecast...

Ensemble created using models: arima, sarima, prophet

Jan 2013Jul 2013Jan 2014Jul 2014Jan 2015Jul 2015Jan 2016Jul 2016Jan 2017Jul 201700.2M0.4M0.6M0.8M1M1.2M1.4M

DataHistorical SalesEnsemble ForecastEnsemble Sales ForecastDateSales

Calculating proportions for store-product combinations...

Generating predictions for each store-product combination...

Processing test rows: 100%

 28512/28512 [00:05<00:00, 5245.58it/s]

Submission file generated: sales\_forecast\_submission.csv

Done!