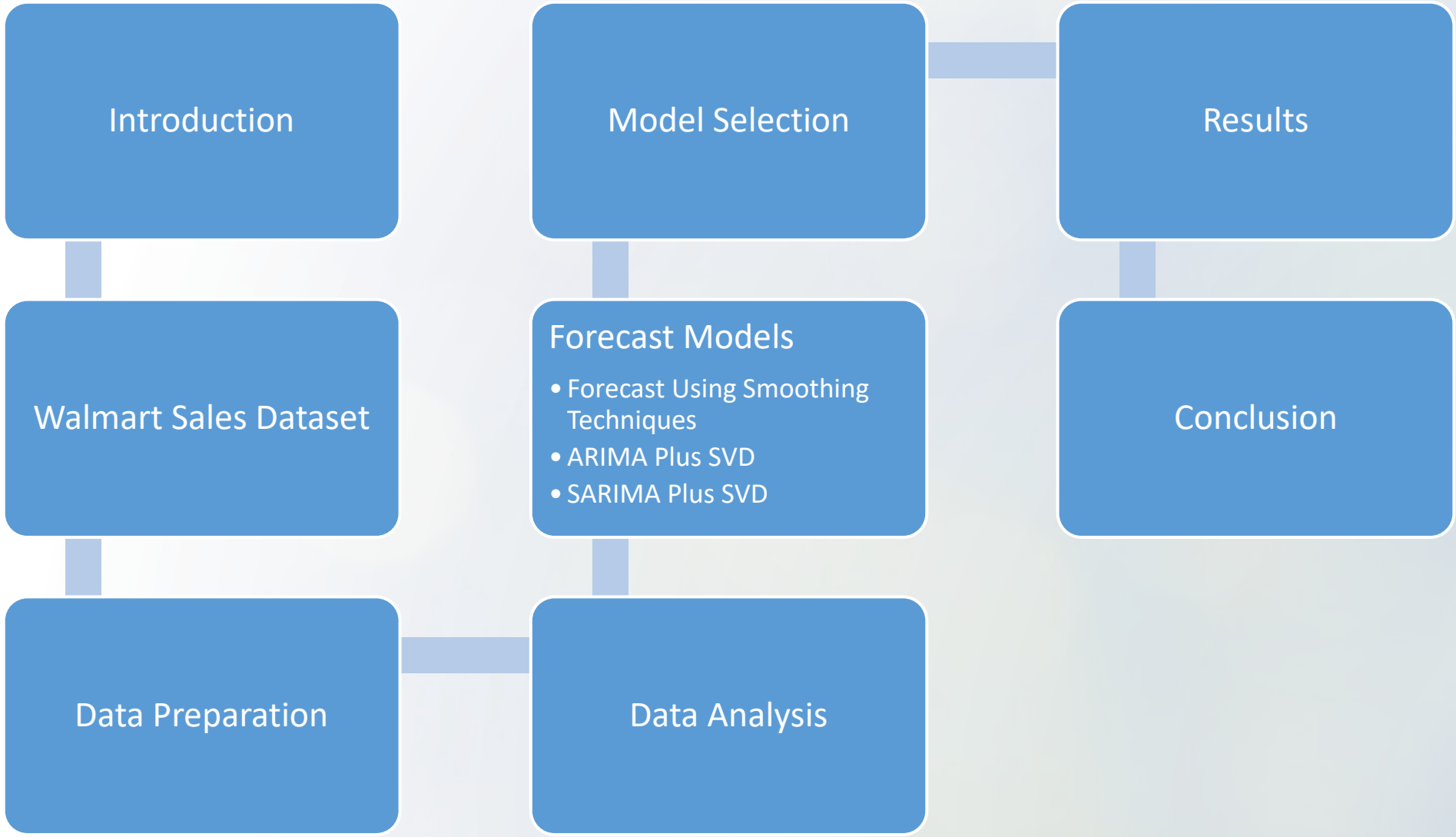
Web Application to Evaluate Statistical Time Series Forecast Models: Application to Walmart Sales

- Author: Ngoc Phan

- App URL: https://sales-forecast.herokuapp.com/
- GitHub Repo: https://github.com/nphan20181/sales-forecast

Agenda


Introduction
Walmart Sales Dataset
Data Preparation
Data Analysis
Forecast Models
• Forecast Using Smoothing Techniques
• ARIMA Plus SVD
• SARIMA Plus SVD
Model Selection
Results
Conclusion

Introduction

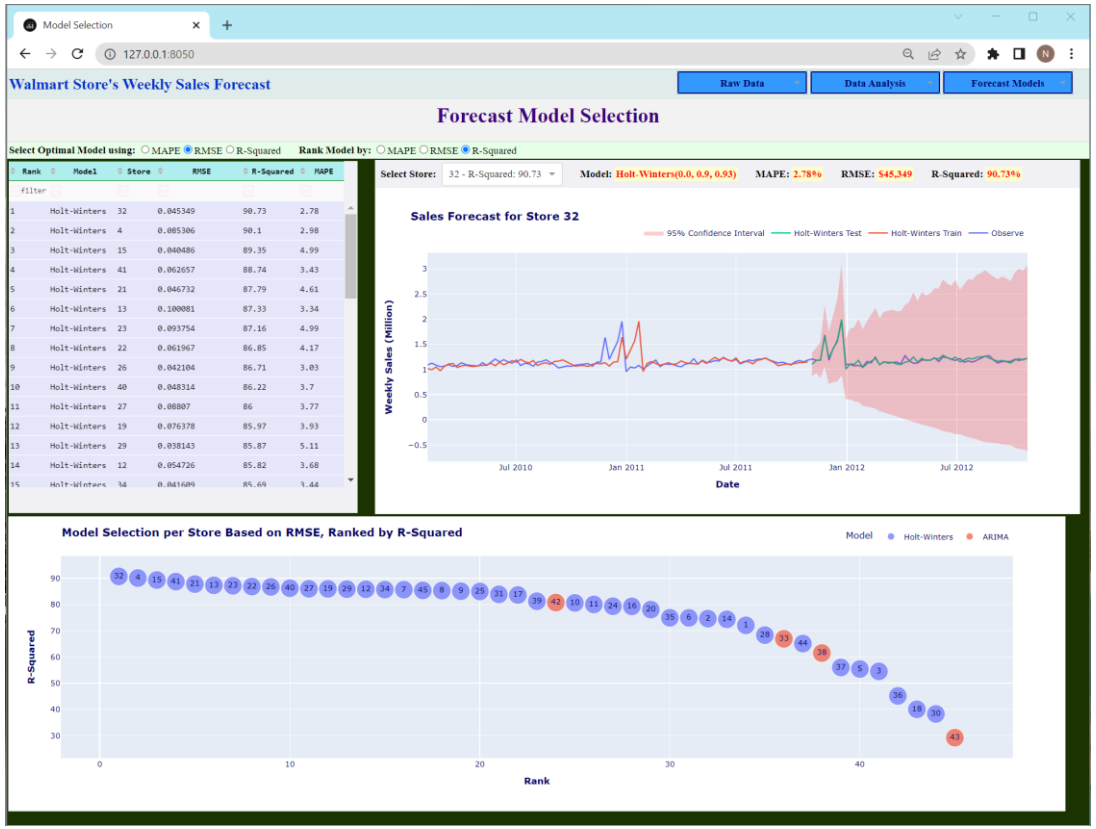
Problem Identification | Proposed Solution | Schematic Diagram

Problem Identification

- Lacking a robust sales forecasting system could potentially attribute to one of the followings:
  - "*93% of sales leaders are unable to forecast revenue within 5 percent, even with two weeks left in the quarter*" (*dooly.ai*).
  - "*By 2025, over 90% of B2B enterprise sales organizations will continue to rely on intuition instead of advanced data analytics, resulting in inaccurate forecasts, sales pipelines, and quota attainment*" (*dooly.ai*).
  - "*Fewer than 20% of sales organizations have forecast accuracy of 75% or greater*" (*InsightSquared*).

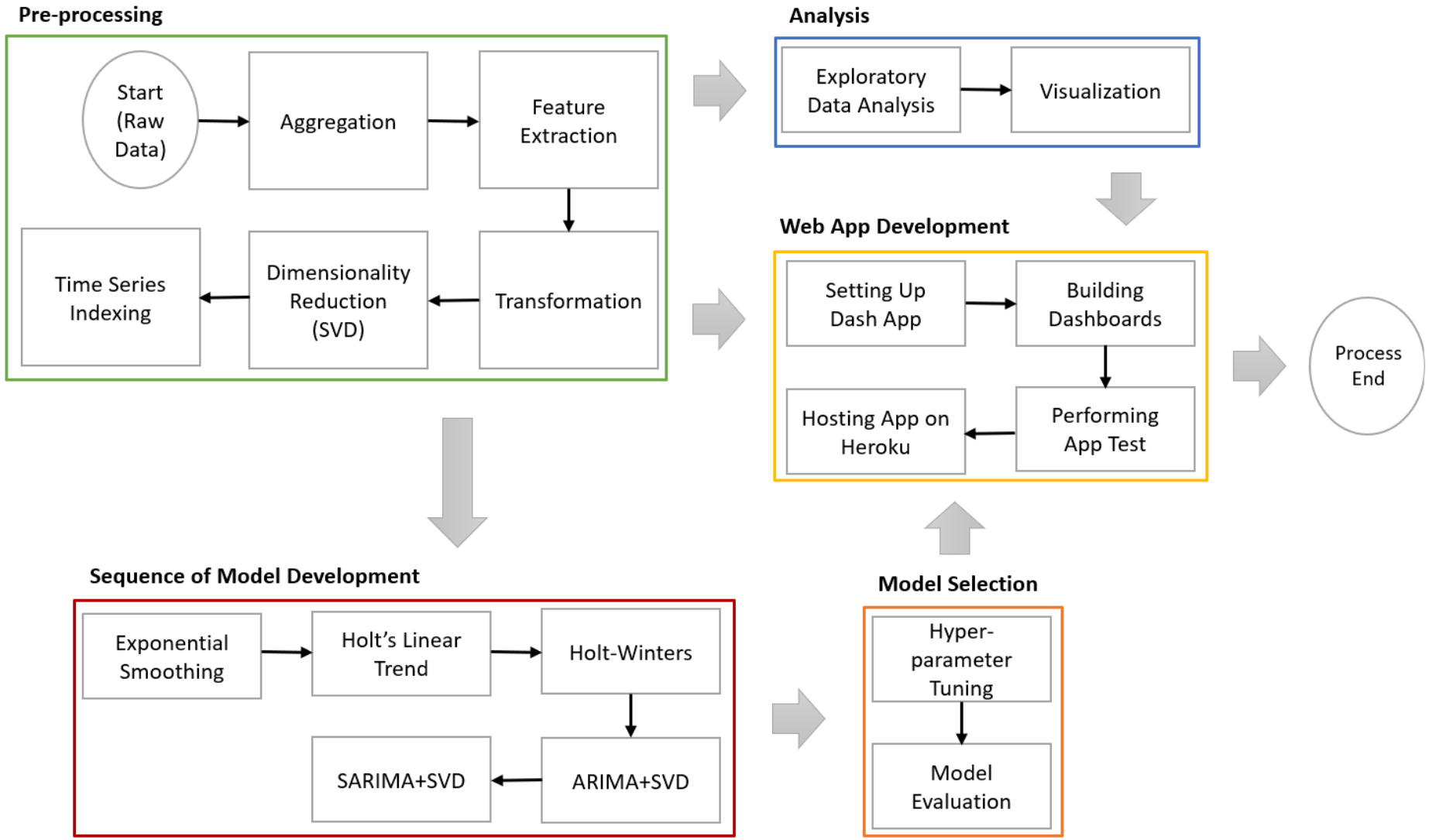
Proposed Solution

- A **sales forecast application** that provides insight into historical data and forecast weekly sales of *one-year* in advance.
  - Enable businesses to set realistic sales target.
  - Help businesses to allocate resources more effectively.
  - Allow businesses to adjust their strategies to meet changing market conditions.



App URL: https://sales-forecast.herokuapp.com

Schematic Diagram of Sales Forecast Application



Walmart Sales Dataset

Train | Stores | Features

Raw Data: train.csv

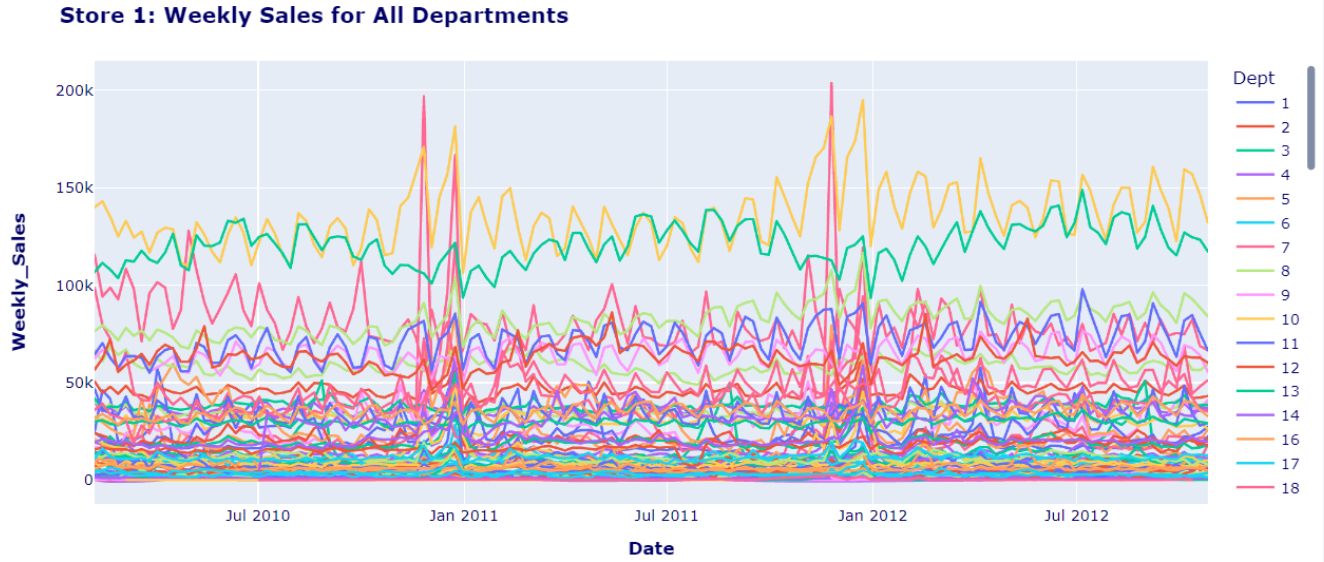
Summary Statistics of Weekly Sales 421,570 observations **Mean:** $15,981.26 **Median:** $7,612.03 **Standard Deviation:** $22,711.18 **Min:** $-4,988.94 **Max:** $693,099.36

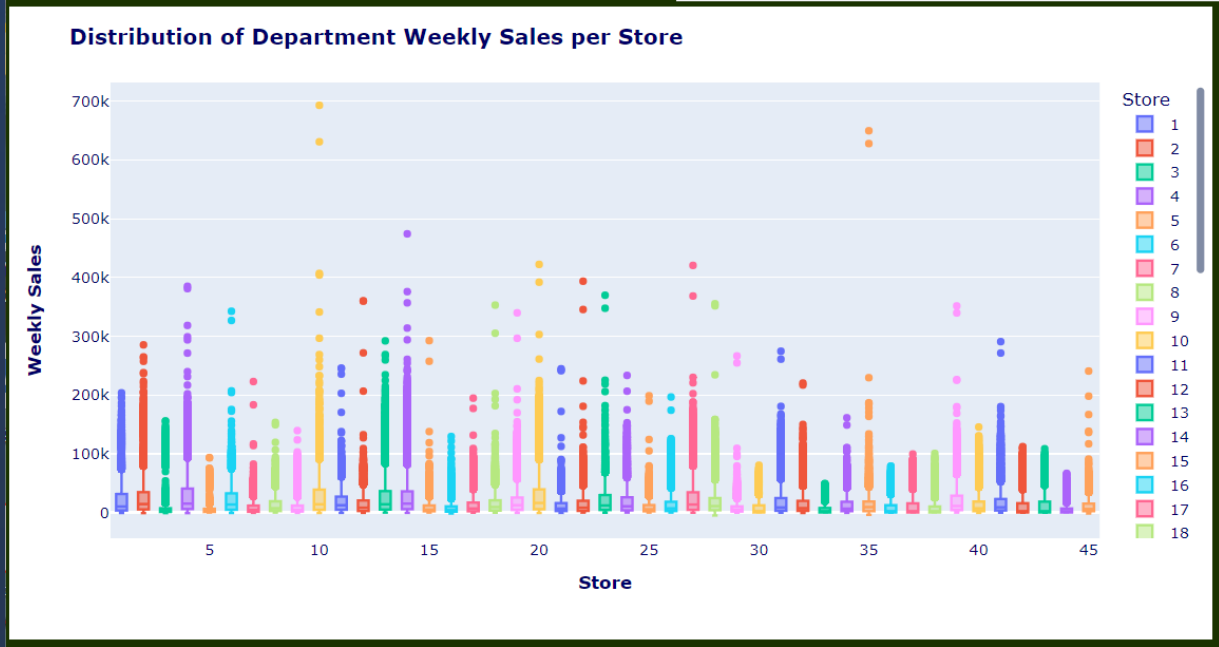
Store: 1 **Department:** All **Observations:** 10,244

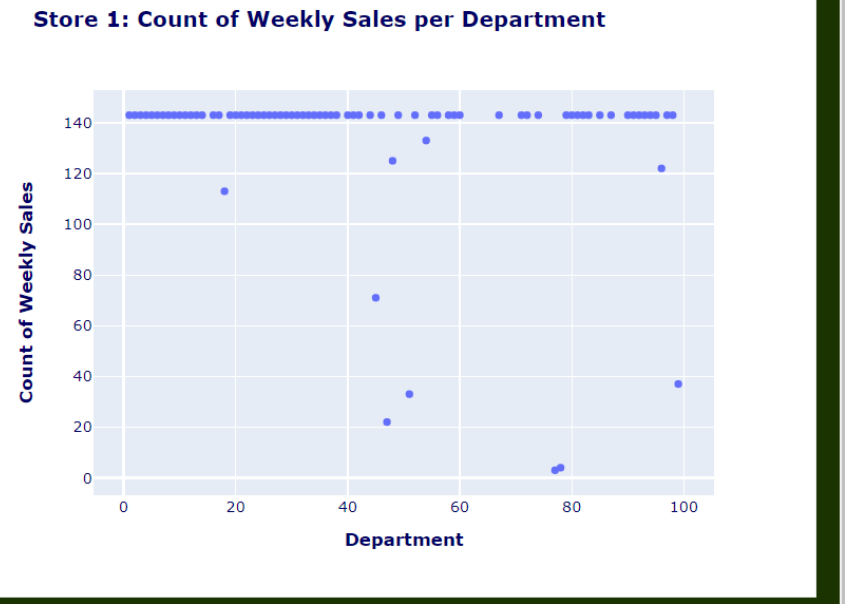
Store	Dept	Date	Weekly_Sales	IsHoliday
1	1	2010-02-05	24924.5	false
1	1	2010-02-12	46039.49	true
1	1	2010-02-19	41595.55	false
1	1	2010-02-26	19403.54	false
1	1	2010-03-05	21827.9	false
1	1	2010-03-12	21043.39	false
1	1	2010-03-19	22136.64	false
1	1	2010-03-26	26229.21	false
1	1	2010-04-02	57258.43	false
1	1	2010-04-09	42960.91	false
1	1	2010-04-16	17596.96	false
1	1	2010-04-23	16145.35	false
1	1	2010-04-30	16555.11	false
1	1	2010-05-07	17413.94	false
1	1	2010-05-14	18926.74	false

« ‹ 1 / 11 › »

Select Store: 1 Select Department: All Number of Departments: 77

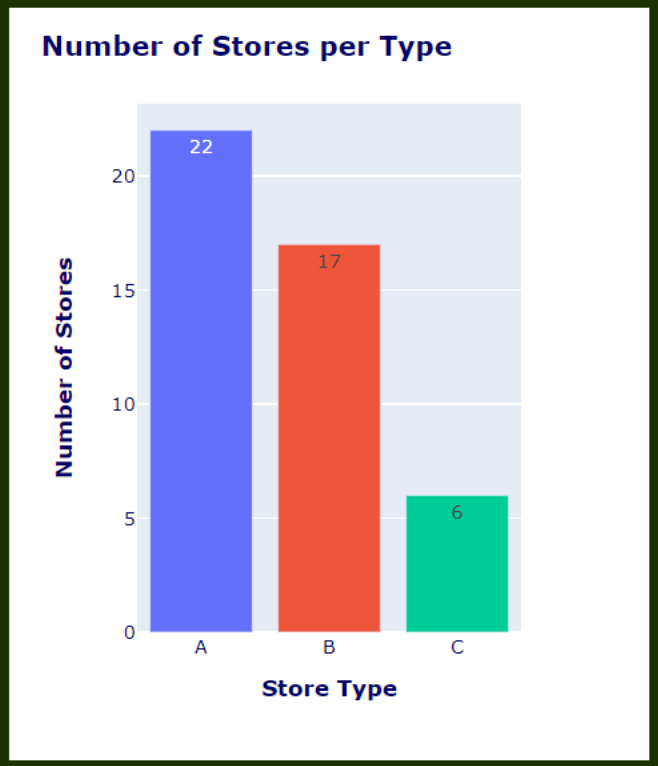


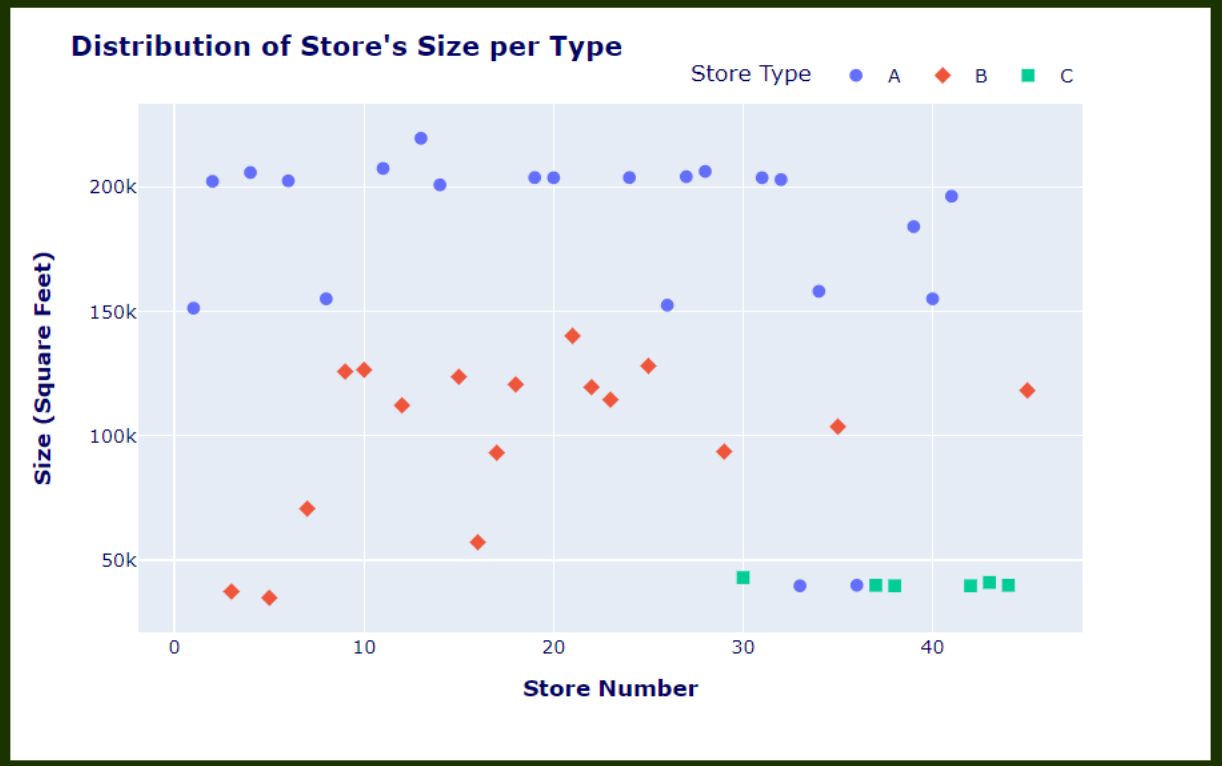




Raw Data: stores.csv

Store	Type	Size
1	A	151315
2	A	202307
3	B	37392
4	A	205863
5	B	34875
6	A	202505
7	B	70713
8	A	155078
9	B	125833
10	B	126512
11	A	207499
12	B	112238
13	A	219622
14	A	200898
15	B	123737

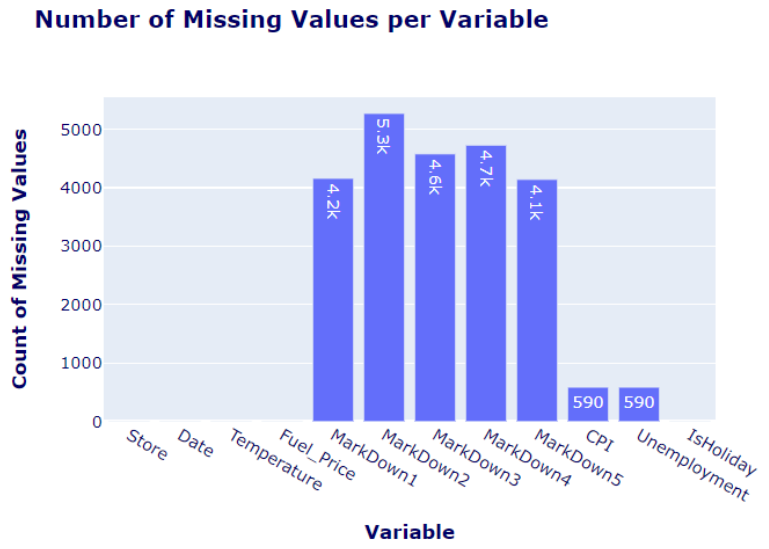




Raw Data: features.csv

Store	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday
1	2010-02-05	42.31	2.572						211.0963582	8.106	false
1	2010-02-12	38.51	2.548						211.2421698	8.106	true
1	2010-02-19	39.93	2.514						211.2891429	8.106	false
1	2010-02-26	46.63	2.561						211.3196429	8.106	false
1	2010-03-05	46.5	2.625						211.3501429	8.106	false
1	2010-03-12	57.79	2.667						211.3806429	8.106	false
1	2010-03-19	54.58	2.72						211.215635	8.106	false
1	2010-03-26	51.45	2.732						211.0180424	8.106	false
1	2010-04-02	62.27	2.719						210.8204499	7.808	false
1	2010-04-09	65.86	2.77						210.6228574	7.808	false
1	2010-04-16	66.32	2.808						210.4887	7.808	false
1	2010-04-23	64.84	2.795						210.4391228	7.808	false
1	2010-04-30	67.41	2.78						210.3895456	7.808	false
1	2010-05-07	72.55	2.835						210.3399684	7.808	false
1	2010-05-14	74.78	2.854						210.3374261	7.808	false

« ‹ 1 / 9 › »

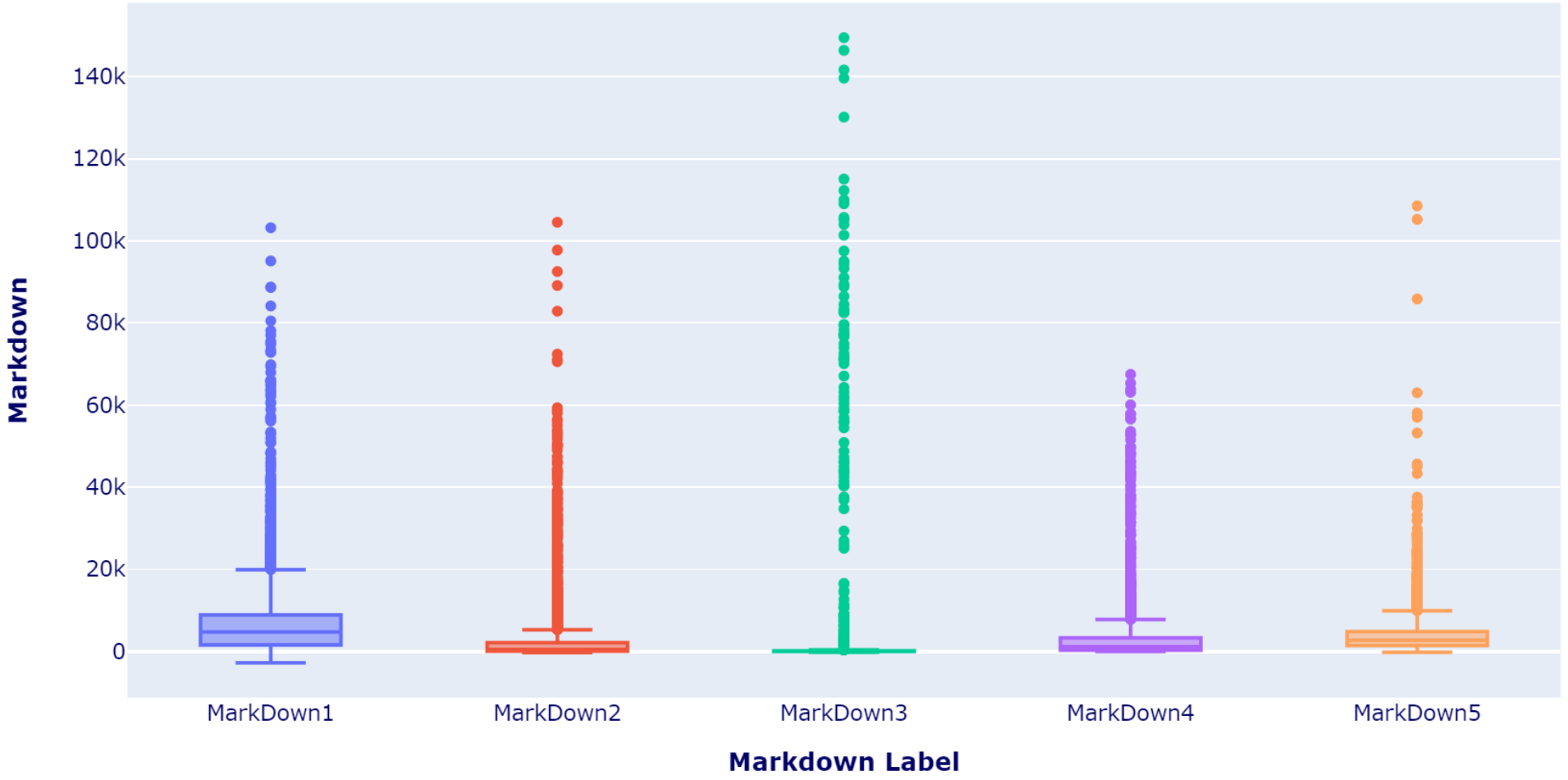


Statistic	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment
count	8190	8190	4032	2921	3613	3464	4050	7605	7605
mean	59.356	3.406	7032.372	3384.177	1760.1	3292.936	4132.216	172.461	7.827
std	18.679	0.431	9262.747	8793.583	11276.462	6792.33	13086.69	39.738	1.877
min	-7.29	2.472	-2781.45	-265.76	-179.26	0.22	-185.17	126.064	3.684
25%	45.902	3.041	1577.532	68.88	6.6	304.688	1440.827	132.365	6.634
50%	60.71	3.513	4743.58	364.57	36.26	1176.425	2727.135	182.764	7.806
75%	73.88	3.743	8923.31	2153.35	163.15	3310.008	4832.555	213.932	8.567
max	101.95	4.468	103184.98	104519.54	149483.31	67474.85	771448.1	228.976	14.313

Distribution of Markdowns (Excluding Values > $700,000)

Markdowns

- available only *after November 2nd, 2011* for particular weeks
- not given by every store's department



Data Preparation

Aggregation & Scaling | Feature Extraction | Transformation |
Dimensionality Reduction | Time Series Indexing

Aggregation & Scaling

- Sum the weekly sales across *all* departments for each store at each given date.
- Divide the aggregated weekly sales by one million.

Date	Store	Department	Weekly Sales
2010-02-05	1	1	24,924.50
2010-02-05	1	2	50,605.27
2010-02-05	1	3	13,740.12
2010-02-05	1	...	...

Date	Store	Weekly Sales (Million)
2010-02-05	1	1.643691
2010-02-12	1	1.641957
...	1	...
2012-10-26	1	1.493660

Feature Extraction

- Extract *week number* from **sales date**
- Scale *week number* so the new value have a range between 0 and 1
  - Scaled_Week = Week / 52
- Create *dummy variables* for special holidays: Super Bowl, Labor Day, Thanksgiving, Before Christmas, and Christmas

Date
2010-02-05
2010-02-12
2010-02-19
…

Date	Week	Scaled_Week	Supper Bowl	Labor Day	Thanksgiving	Before Christmas	Christmas
2010-02-05	5	1.6436909	0	0	0	0	0
2010-02-12	6	1.6419574	1	0	0	0	0
2010-02-19	7	1.6119682	0	0	0	0	0
…	…	…	…	…	…	…	…

Transformation

- Apply log transformation to ***Weekly Sales (Million)***

Date	Store	Weekly Sales (Million)
2010-02-05	1	1.643691
2010-02-12	1	1.641957
…	1	…
2012-10-26	1	1.493660

Date	Store	Weekly Sales (Million)	Log Of Weekly Sales (Million)
2010-02-05	1	1.643691	0.49694426192494
2010-02-12	1	1.641957	0.495889091084086
…	1	…	…
2012-10-26	1	1.493660	0.401229309767411

Dimensionality Reduction: Singular Value Decomposition (SVD)

Six features

- *Scaled_Week, Super Bowl, Labor Day, Thanksgiving, Before Christmas,* and *Christmas*

Algorithm: searching for number of optimal SVD components

- Iterate through store 1 to 45
  - Retrieve sales data for the given store
  - Iterate through *number of components*, ***n***, from 1 to 4
    - Reduce six features to *n components* using **TruncatedSVD** from Python's library, *sklearn*
    - Fit an ARIMA(1,1,1) + SVD(n) model using ***Log of Weekly Sales (Million)*** as target variable
    - Save model's results
  - Select a model that has a lowest ***Akaike Information Criterion (AIC)*** score
  - Save *number of SVD components* for respective store

Time Series Indexing

- Sort sales data for each store by ***Date*** in ascending order
- Set a time period of *one* for the *first* weekly sales, a time period of *two* for the *second* weekly sales, and so on.

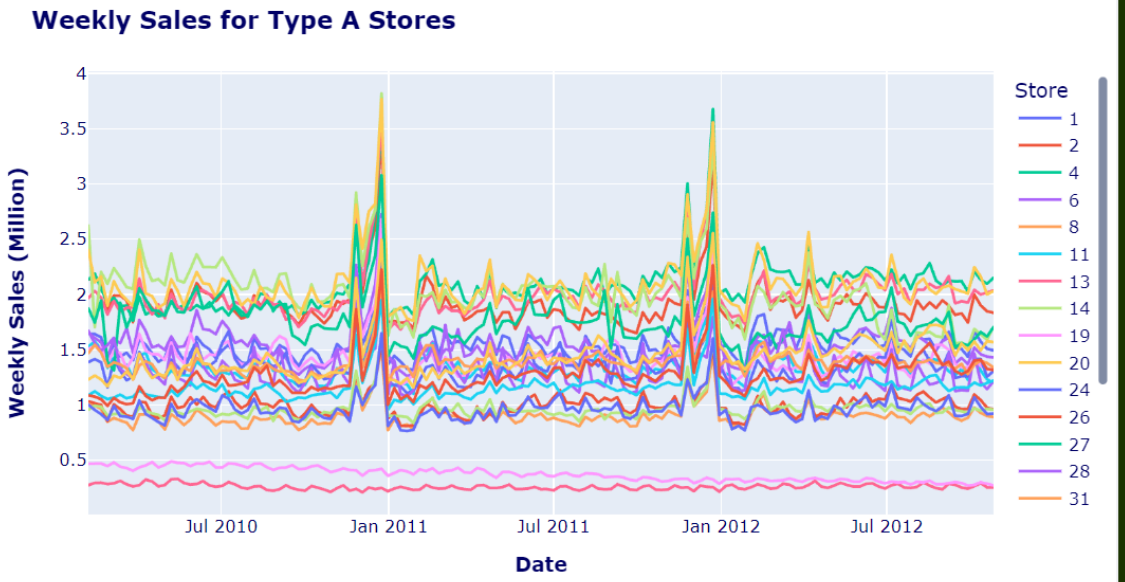
Time Period	Date	Store	Weekly Sales (Million)	Log Of Weekly Sales (Million)
1	2010-02-05	1	1.643691	0.49694426192494
2	2010-02-12	1	1.641957	0.495889091084086
...	...	1	...	...
143	2012-10-26	1	1.493660	0.401229309767411

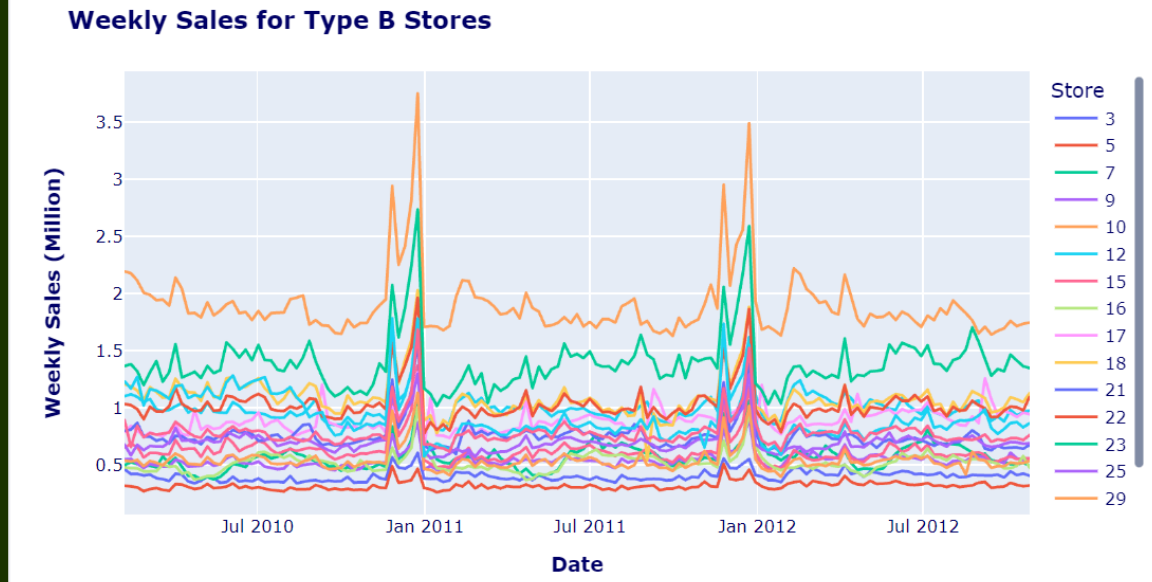
Data Analysis

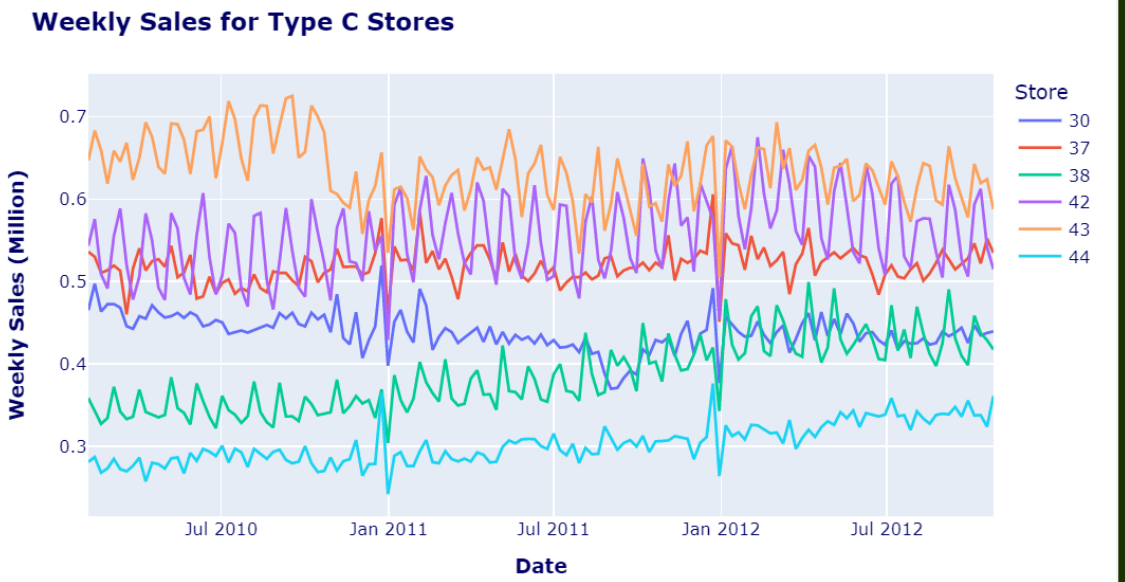
Weekly Sales per Store Type | Weekly Sales Correlation per Store

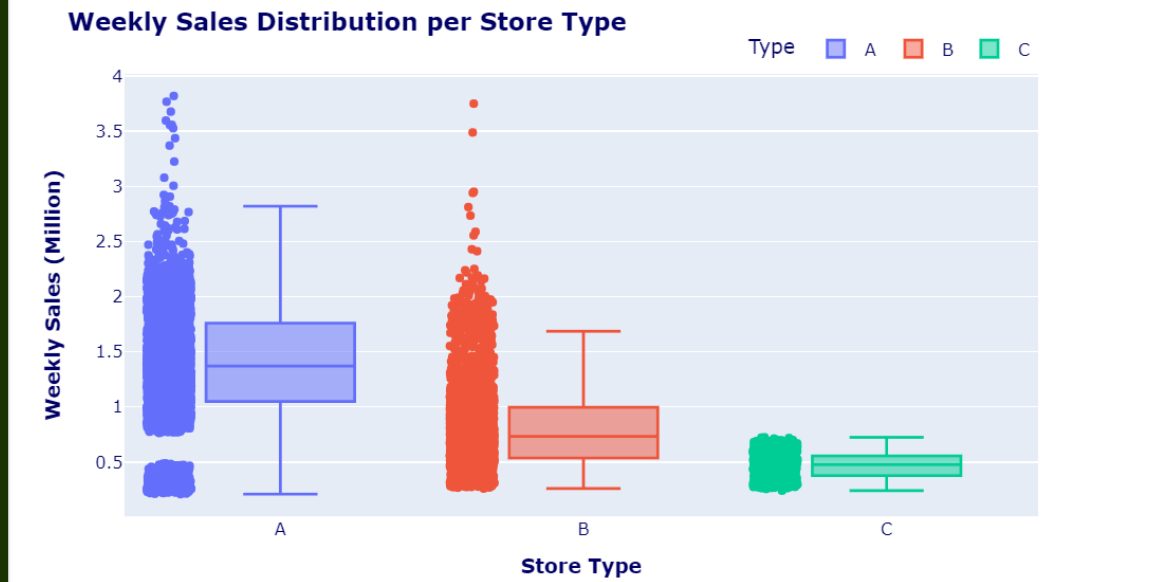
Walmart Store's Weekly Sales Analysis

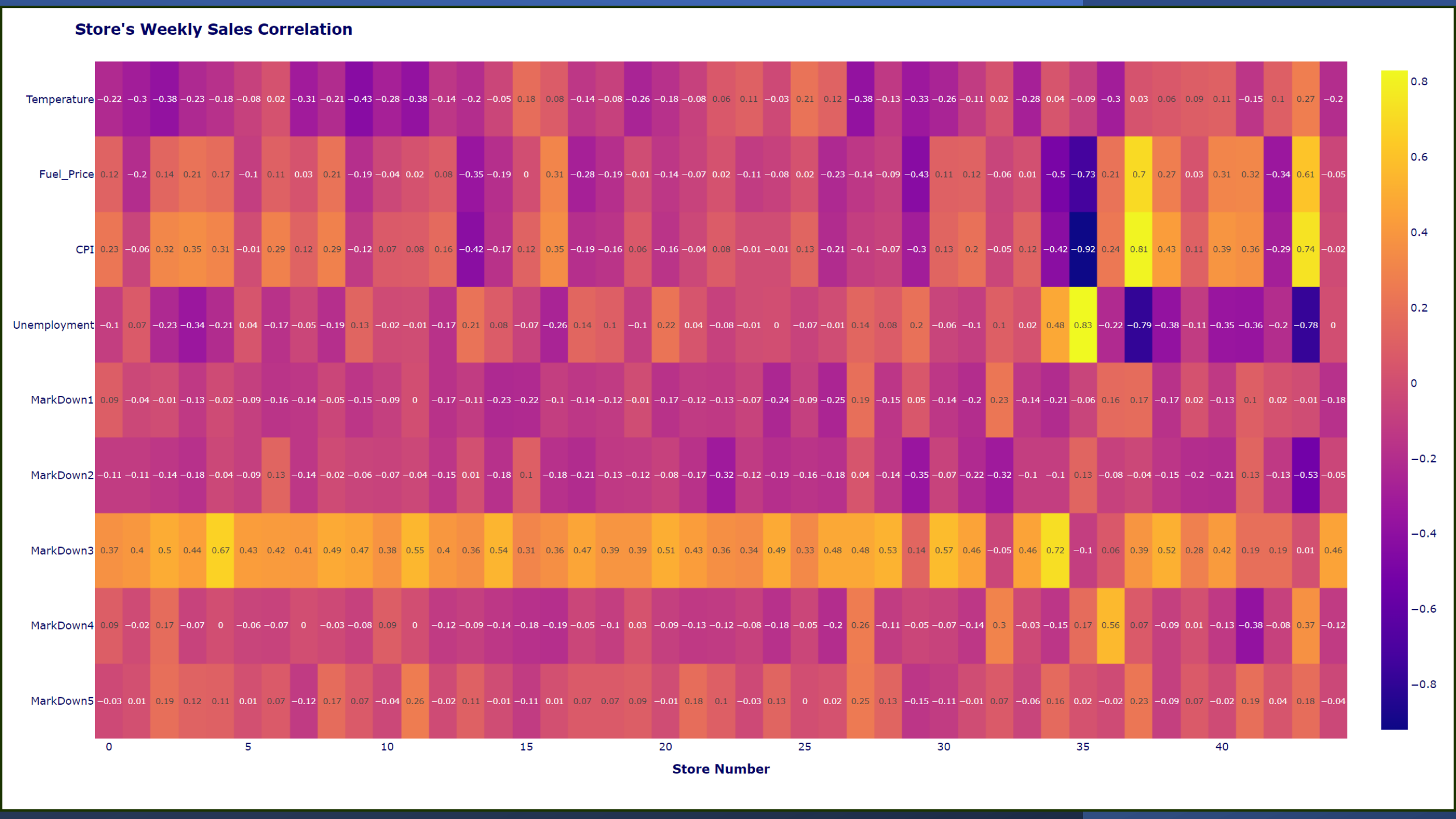
Summary Statistics of Store's Weekly Sales 6,435 observations Mean: $1,046,964.88 Median: $960,746.04 Standard Deviation: $564,366.62 Min: $209,986.25 Max: $3,818,686.45










Store's Weekly Sales Correlation
Temperature −0.22 −0.3 −0.38 −0.23 −0.18 −0.08 0.02 −0.31 −0.21 −0.43 −0.28 −0.38 −0.14 −0.2 −0.05 0.18 0.08 −0.14 −0.08 −0.26 −0.18 −0.08 0.06 0.11 −0.03 0.21 0.12 −0.38 −0.13 −0.33 −0.26 −0.11 0.02 −0.28 0.04 −0.09 −0.3 0.03 0.06 0.09 0.11 −0.15 0.1 0.27 −0.2
Fuel_Price 0.12 −0.2 0.14 0.21 0.17 −0.1 0.11 0.03 0.21 −0.19 −0.04 0.02 0.08 −0.35 −0.19 0 0.31 −0.28 −0.19 −0.01 −0.14 −0.07 0.02 −0.11 −0.08 0.02 −0.23 −0.14 −0.09 −0.43 0.11 0.12 −0.06 0.01 −0.5 −0.73 0.21 0.7 0.27 0.03 0.31 0.32 −0.34 0.61 −0.05
CPI 0.23 −0.06 0.32 0.35 0.31 −0.01 0.29 0.12 0.29 −0.12 0.07 0.08 0.16 −0.42 −0.17 0.12 0.35 −0.19 −0.16 0.06 −0.16 −0.04 0.08 −0.01 −0.01 0.13 −0.21 −0.1 −0.07 −0.3 0.13 0.2 −0.05 0.12 −0.42 −0.92 0.24 0.81 0.43 0.11 0.39 0.36 −0.29 0.74 −0.02
Unemployment −0.1 0.07 −0.23 −0.34 −0.21 0.04 −0.17 −0.05 −0.19 0.13 −0.02 −0.01 −0.17 0.21 0.08 −0.07 −0.26 0.14 0.1 −0.1 0.22 0.04 −0.08 −0.01 0 −0.07 −0.01 0.14 0.08 0.2 −0.06 −0.1 0.1 0.02 0.48 0.83 −0.22 −0.79 −0.38 −0.11 −0.35 −0.36 −0.2 −0.78 0
MarkDown1 0.09 −0.04 −0.01 −0.13 −0.02 −0.09 −0.16 −0.14 −0.05 −0.15 −0.09 0 −0.17 −0.11 −0.23 −0.22 −0.1 −0.14 −0.12 −0.01 −0.17 −0.12 −0.13 −0.07 −0.24 −0.09 −0.25 0.19 −0.15 0.05 −0.14 −0.2 0.23 −0.14 −0.21 −0.06 0.16 0.17 −0.17 0.02 −0.13 0.1 0.02 −0.01 −0.18
MarkDown2 −0.11 −0.11 −0.14 −0.18 −0.04 −0.09 0.13 −0.14 −0.02 −0.06 −0.07 −0.04 −0.15 0.01 −0.18 0.1 −0.18 −0.21 −0.13 −0.12 −0.08 −0.17 −0.32 −0.12 −0.19 −0.16 −0.18 0.04 −0.14 −0.35 −0.07 −0.22 −0.32 −0.1 −0.1 0.13 −0.08 −0.04 −0.15 −0.2 −0.21 0.13 −0.13 −0.53 −0.05
MarkDown3 0.37 0.4 0.5 0.44 0.67 0.43 0.42 0.41 0.49 0.47 0.38 0.55 0.4 0.36 0.54 0.31 0.36 0.47 0.39 0.39 0.51 0.43 0.36 0.34 0.49 0.33 0.48 0.48 0.53 0.14 0.57 0.46 −0.05 0.46 0.72 −0.1 0.06 0.39 0.52 0.28 0.42 0.19 0.19 0.01 0.46
MarkDown4 0.09 −0.02 0.17 −0.07 0 −0.06 −0.07 0 −0.03 −0.08 0.09 0 −0.12 −0.09 −0.14 −0.18 −0.19 −0.05 −0.1 0.03 −0.09 −0.13 −0.12 −0.08 −0.18 −0.05 −0.2 0.26 −0.11 −0.05 −0.07 −0.14 0.3 −0.03 −0.15 0.17 0.56 0.07 −0.09 0.01 −0.13 −0.38 −0.08 0.37 −0.12
MarkDown5 −0.03 0.01 0.19 0.12 0.11 0.01 0.07 −0.12 0.17 0.07 −0.04 0.26 −0.02 0.11 −0.01 −0.11 0.01 0.07 0.07 0.09 −0.01 0.18 0.1 −0.03 0.13 0 0.02 0.25 0.13 −0.15 −0.11 −0.01 0.07 −0.06 0.16 0.02 −0.02 0.23 −0.09 0.07 −0.02 0.19 0.04 0.18 −0.04
0 5 10 15 20 25 30 35 40
Store Number
0.8 0.6 0.4 0.2 0 −0.2 −0.4 −0.6 −0.8

Forecast Using Smoothing Techniques

Exponential Smoothing | Holt’s Linear Trend | Multiplicative Holt-Winters

Exponential Smoothing

- Weight most recent observation heavier than distant observations.
- The method is appropriate for a time series data that changes slowly over time and exhibits no seasonal pattern and no trend.
- Target variable: ***Weekly Sales (Million)***

Level	$\ell_t = \alpha y_t + (1-\alpha)\ell_{t-1}$
In-sample Forecast	$\hat{y}_t = \ell_{t-1}$
Out-of-sample Forecast	$\hat{y}_{T+\tau} = \ell_T \quad \text{for} \quad \tau = 1, 2, \cdots$
Standard Error	$s = \sqrt{\frac{SSE}{T-1}}$
95% Prediction Interval	$\left[\ell_T \pm z_{[.025]} s\sqrt{1+(\tau-1)\alpha^2}\right]$

Note:

- $\boldsymbol{\ell_t}$ is the level or mean at time period t.
- Initial level $\boldsymbol{\ell_0} = average\ sales\ at\ week\ 4$
- $\boldsymbol{y_t}$ is the observe value at time period t.
- $\boldsymbol{\alpha}$ is a smoothing constant.
- **SSE** is sum of squared error.
- **T** is the last time period in the series.

Holt's Linear Trend

- Appropriate for a time series data having both level and growth rate that change over time.
- Target variable: ***Weekly Sales (Million)***

Level	$\ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} - b_{t-1})$
Growth Rate	$b_t = \gamma(\ell_t - \ell_{t-1}) + (1-\gamma)b_{t-1}$
In-sample Forecast	$\hat{y}_t = \ell_{t-1} + b_{t-1}$
Out-of-sample Forecast	$\hat{y}_{T+\tau} = \ell_T + \tau b_T \quad \text{for} \quad \tau = 1, 2, \cdots$
Standard Error	$s = \sqrt{\frac{SSE}{T-2}}$
95% Prediction Interval If $\tau = 1$: $[(\ell_T + b_T) \pm z_{[.025]}s]$ If $\tau \geq 2$: $\left[(\ell_T + \tau b_T) \pm z_{[.025]}s\sqrt{1 + \sum_{j=1}^{\tau-1}\alpha^2(1+j\gamma)^2}\right]$	

- Algorithm: computing initial level and growth rate
  - Fit a least squares regression line using the ***first 52 weeks of data***
  - Set initial level = intercept
  - Set initial growth rate = coefficient

$$y_t = \beta_0 + \beta_1 \times t$$

Initial level ℓ_0 ← β_0 ; β_1 → Initial growth rate b_0

Multiplicative Holt-Winters

- Appropriate for a time series data that exhibits a linear trend and *changing* level, growth rate and seasonal pattern.
- Target variable: ***Weekly Sales (Million)***

Level	$\ell_t = \alpha(y_t/sn_{t-L}) + (1-\alpha)(\ell_{t-1} + b_{t-1})$
Growth Rate	$b_t = \gamma(\ell_t - \ell_{t-1}) + (1-\gamma)(b_{t-1})$
Seasonal Factor	$sn_t = \delta(y_t/\ell_t) + (1-\delta)(sn_{t-L})$
In-sample Forecast	$\hat{y}_t = (\ell_{t-1} + b_{t-1})sn_{t-L}$
Out-of-sample Forecast	$\hat{y}_{T+\tau} = (\ell_T + \tau b_T)sn_{T+\tau-L}$ for $\tau = 1, 2, \cdots$
Relative Standard Error	$s_r = \sqrt{\frac{\sum_{t=1}^{T}\left[\frac{y_t - (\ell_{t-1} + b_{t-1})sn_{t-L}}{(\ell_{t-1} + b_{t-1})sn_{t-L}}\right]^2}{T-3}}$
95% Prediction Interval	$\left[\hat{y}_{T+\tau}(T) \pm z_{[.025]}s_r(\sqrt{c_\tau})(sn_{T+\tau-L})\right]$ • If $\tau = 1$ then $c_1 = (\ell_T + b_T)^2$ • If $2 \le \tau \le L$ then $c_\tau = \sum_{j=1}^{\tau-1}\alpha^2(1 + [\tau - j]\gamma)^2(\ell_T + jb_T)^2 + (\ell_T + \tau b_T)^2$

Multiplicative Holt-Winters (cont.)

- Steps for computing initial values
  1) Fit a least squares trend line on the entire training data.
     - Set ***initial level*** = intercept.
     - Set ***initial growth rate*** = coefficient.
  2) Compute a regression estimate, $Trend_t$, at each time period t using the least squares trend line in step 1.
  3) Detrend the data at each time period t:
$$D_t = y_t / Trend_t$$
  4) Average detrended values for each of the 52 weeks. As a result, we have a series of average seasonal values for week 1 to 52, denoted as $\bar{S}_1, \bar{S}_2, \cdots, \bar{S}_{52}$.
  5) Compute the seasonal correct factor CF. Note that L = 52 (number of weeks in a year)
$$CF = \frac{L}{\sum_{i=1}^{L} \bar{S}_i}$$
  6) Compute ***initial seasonal factors***, sn_{i-L}, at time periods $i - L$ where $i = 1, 2, \cdots, L$.
$$sn_{i-L} = \bar{S}_i \times CF$$

Autoregressive Integrated Moving Average (ARIMA) Plus SVD

Moving Average | Autoregressive | ARIMA | Process Flow Diagram |
Augmented Dickey-Fuller Test for Stationarity |
Transformation for Non-Stationary Series |
Ljung Box Test for Uncorrelated Residuals

Moving Average Process

- A moving average model is denoted as **MA(q)** where $\boldsymbol{q}$ is the order.
- The model expresses the present value $\boldsymbol{y_t}$ as a linear combination of
  - the mean of the series $\boldsymbol{\mu}$
  - the present error term $\boldsymbol{\varepsilon_t}$
  - the past error terms $\boldsymbol{\varepsilon_{t-i}}$

$$y_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

- The magnitude of the impact of past errors on the present value is quantified using a coefficient denoted as $\boldsymbol{\theta_i}$.
- The error terms are assumed to be mutually independent and normally distributed, just like white noise.

Autoregressive Process

- An autoregressive model is denoted as an **AR(p)** process, where $\boldsymbol{p}$ is the order.
- The model expresses the present value $\boldsymbol{y_t}$ is a linear combination of
  - a constant $\boldsymbol{C}$
  - the present error term $\boldsymbol{\varepsilon_t}$, which is also *white noise*
  - the past values of the series $\boldsymbol{y_{t-j}}$.

$$y_t = C + \varepsilon_t + \sum_{j=1}^{p} \phi_j y_{t-j}$$

- The magnitude of the influence of the past values on the present value is denoted as $\boldsymbol{\phi_j}$, which represents the coefficients of the **AR(p)** model.

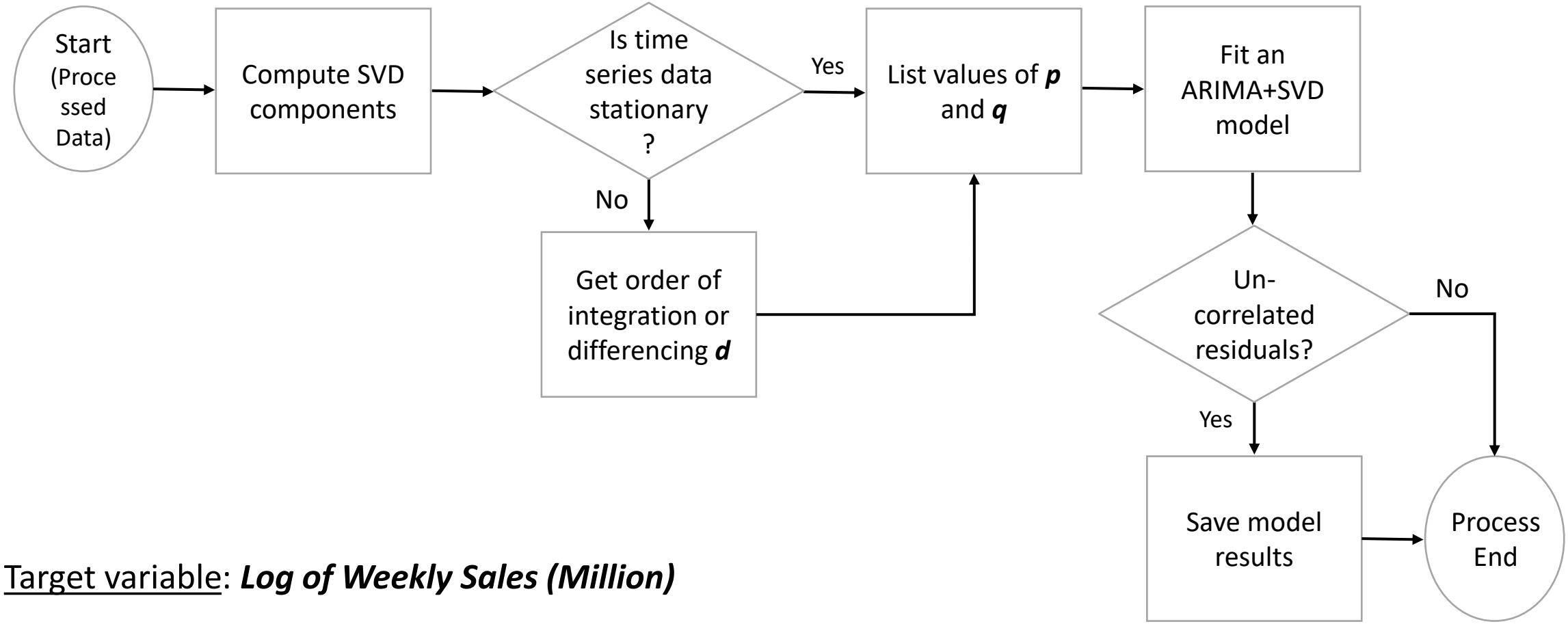
$ARIMA(p, d, q)$

- ARIMA process is the combination of an autoregressive process ***AR(p)***, integration ***I(d)***, and moving average process ***MA(q)***.
- Instead of using the original series y_t, ARIMA process uses the differenced series y'_t.
- The mathematical expression of the **ARIMA(*p,d,q*)+SVD(*n*)** process states that the present value of the differenced series y'_t is a linear combination of:
  - a constant C
  - the mean of the differenced series μ
  - a current error term ε_t
  - past error terms $\theta_i \varepsilon'_{t-i}$
  - past values of the differenced series $\phi_j y'_{t-j}$
  - SVD components $\beta_k x_{k,t}$, where ***n*** is the number of components

Note: ***coefficients*** are computed using ***maximum likelihood estimates***.

$$y'_t = C + \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon'_{t-i} + \sum_{j=1}^{p} \phi_j y'_{t-j} + \sum_{k=1}^{n} \beta_k x_{k,t}$$

ARIMA Process Flow Diagram



Target variable: ***Log of Weekly Sales (Million)***

Augmented Dickey-Fuller (ADF) Test for Stationary Time Series Data

A ***stationary time series*** has a constant mean, variance, and autocorrelation, and these properties are independent of time.

ADF Hypotheses

- Null hypothesis: there is a *unit root* present in a time series.
- Alternative hypothesis: there is no *unit root*, time series is stationary.

Result of ADF test

- ***ADF statistic***, which is a negative number
  - The more negative it is, the stronger the rejection of the null hypothesis.
- ***p-value***
  - If it is less than 0.05, we can reject the null hypothesis and say the series is stationary.

Transformation for Non-Stationary Series

- Applying a ***log*** function to the series can stabilize its variance.
- Differencing
  - ***Differencing*** involves calculating the series of changes from one timestep to another.

$$y'_t = y_t - y_{t-1}$$

  - This transformation helps stabilize the mean, which in turn removes or reduces the trend and seasonality effects.
  - Taking the difference makes us lose one data point, because at the initial point in time, we cannot take the difference with its previous step, since $t = -1$ does not exist.

Ljung Box Test for Un-correlated Residuals

The test determines whether or not the errors are *independent and identically distributed* (***iid***).

It is a test of ***lack of fit***.

- If the autocorrelations of the residuals are very small, the model doesn't show a ***significant lack of fit***.

Test Hypotheses

- Null hypothesis: the model *does not* show a lack of fit.
- Alternative hypothesis: the model *does* show a lack of fit.

Accept null hypothesis if p-value > 0.5

Seasonal Autoregressive Integrated Moving Average (SARIMA) Plus SVD

SARIMA | Process Flow Diagram

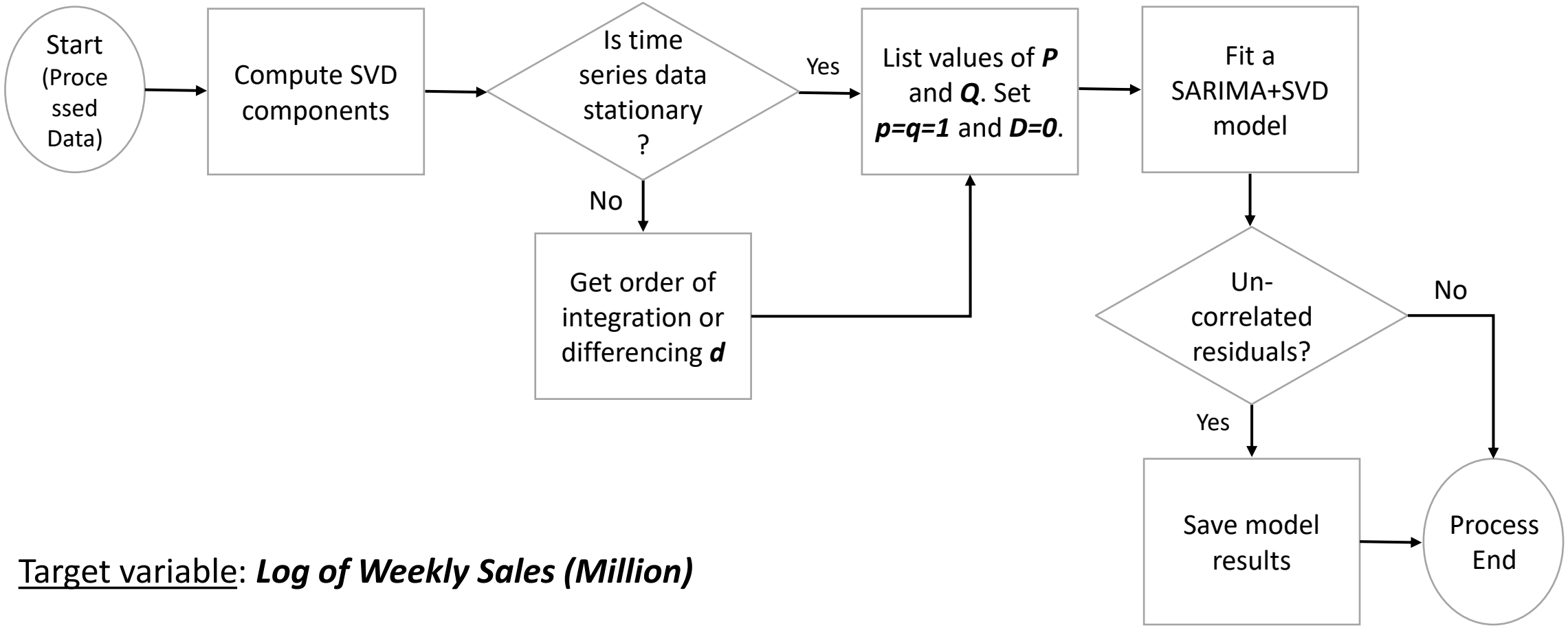
$SARIMA(p,d,q)(P,D,Q)_m$

- Expands on ARIMA(p,d,q) model
  - The first three parameters have the same meaning as in ***ARIMA(p,d,q)*** model.
  - ***m*** is the number of observations per cycle (*e.g.,* number of weeks per year)
  - ***P*** is the order of the ***seasonal AR(P)*** process
    - If P = 2, we are including two past values of the series at a lag that is a multiple of m.
    - For m=52, we include the values at $\boldsymbol{y_{t-1\times 52}}$ and $\boldsymbol{y_{t-2\times 52}}$.
  - ***D*** is the ***seasonal order of integration***
    - If D = 1, this means that a seasonal difference makes the series stationary.

$$y'_t = y_t - y_{t-m}$$

  - ***Q*** is the order of the ***seasonal MA(Q)*** process
    - If Q = 2, we include past error terms at lags that are a multiple of m. Therefore, we will include the errors $\boldsymbol{\varepsilon_{t-1\times m}}$ and $\boldsymbol{\varepsilon_{t-2\times m}}$.

SARIMA Process Flow Diagram



Target variable: ***Log of Weekly Sales (Million)***

Model Selection

Evaluation Metrics | Hyperparameter Tuning

Evaluation Metrics

- **Root Mean Squared Error (RMSE)**

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{n}}$$

  - A measure of the *variation* in the distribution of forecast errors.

- **Mean Absolute Percentage Error (MAPE)**

$$MAPE = \frac{\sum_{t=1}^{n}\frac{|y_t - \hat{y}_t|}{y_t}}{n}(100)$$

  - A summary measure for the *center* of the distribution of forecast errors.
  - Allows comparison across different time series with values of different magnitude.
  - Assumes the values of time series are *positive*.

Evaluation Metrics (cont.)

- **Simple Coefficient of Determination, r^2**
  - $Explained\ Variation = \sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2$
  - $Total\ Variation = \sum_{t=1}^{n}(y_t - \bar{y})^2$
  - r^2 is the proportion of the total variation in the *n* observed values of the dependent variable that is *explained* by the simple linear regression model.

$$r^2 = \frac{Explained\ Variation}{Total\ Variation}$$

  - The closer r^2 is to 1, the larger is the proportion of the total variation that is *explained* by the model, and the greater is the utility of the model in predicting actual value.

Evaluation Metrics (cont.)

- **Akaike Information Criterion (AIC)**

$$AIC = 2K - 2ln(L)$$

  - $\boldsymbol{K}$ is the ***number of independent variables*** used to build the model.
  - $\boldsymbol{L}$ is the ***maximum likelihood estimate*** of the model
    - Indicate how well the model reproduces the data.
  - The smaller is the AIC value, the better is the model.

Hyperparameter Tuning

- Train size
  - first 91 observations
- Test size
  - last 52 observations

Exponential Smoothing

- For each store, fit a model on various alpha value ranging from ***0.0*** to ***1.0***

Holt's Linear Trend

- For each store, fit a model on a combination of alpha and gamma values, each ranging from ***0.0*** to ***1.0***

Holt-Winters

- Run ***1,500*** iterations
  - randomly generate a list of *unique* combination of alpha, gamma, and delta values, each ranging from 0.0 to 1.0
- For each store, fit a model on each 1,500 combinations of alpha, gamma, and delta values obtained from previous step.

ARIMA plus SVD

- For each store, fit a model on a combination of ***p*** and ***q*** values, each ranging from ***0*** to ***12***.

SARIMA plus SVD

- For each store, fit a model on a combination of ***P*** and ***Q*** values, each ranging from ***0*** to ***3***.

Results

Model Selection Result | Model Results

Model Selection Result

Model selection based on *lowest RMSE* per store.

Model ranking based on *MAPE* across ***all*** stores.

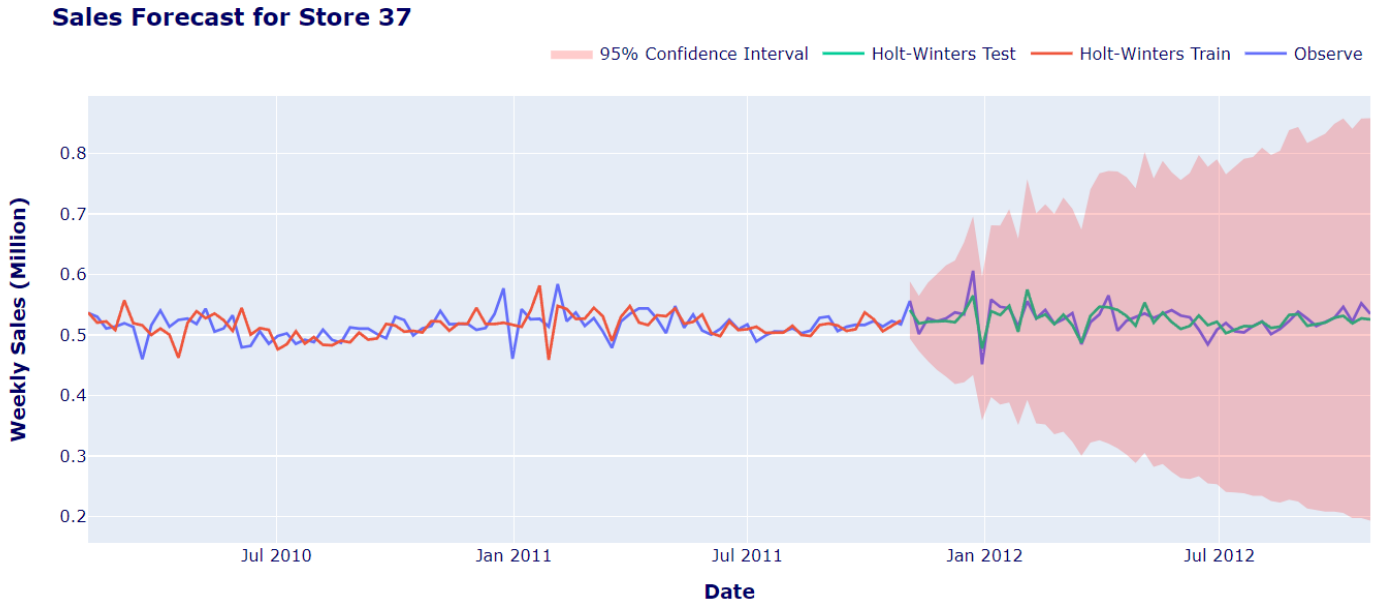
From 45 selected models in making sales forecast

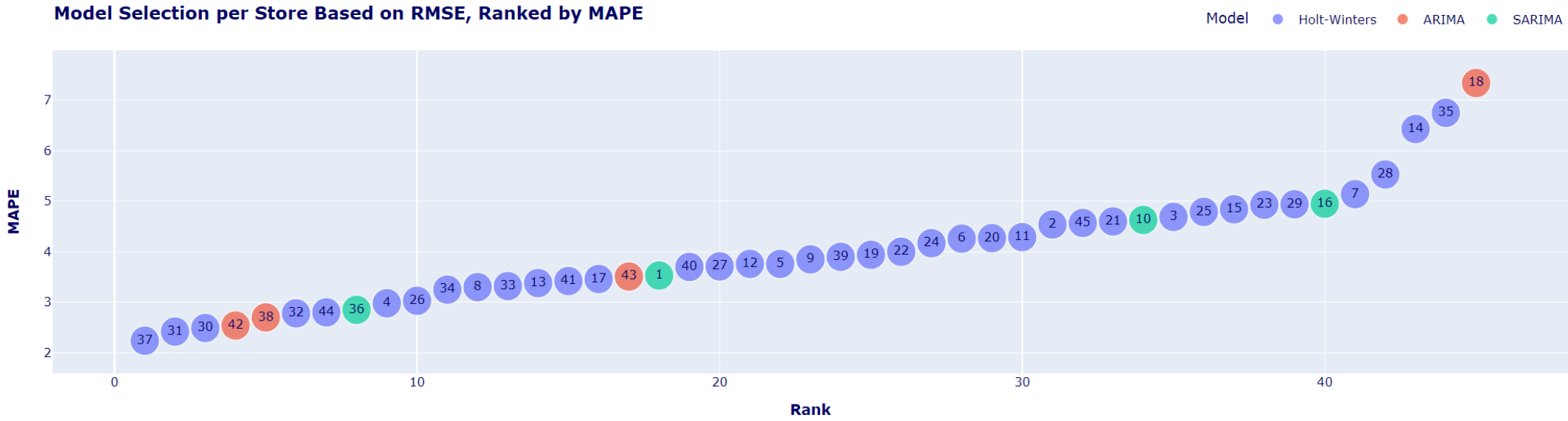
- ***RMSE*** ranges from ***$11,081*** to ***$160,250***
- ***MAPE*** ranges from ***2.24%*** to ***7.34%***
- ***R-Squared*** ranges from ***38.78%*** to ***90.75%***

Select Optimal Model using: ○ MAPE ◉ RMSE ○ R-Squared **Rank Model by:** ◉ MAPE ○ RMSE ○ R-Squared

Rank	Model	Store	RMSE	R-Squared	MAPE
filter					
1	Holt-Winters	37	0.015054	55.97	2.24
2	Holt-Winters	31	0.050715	84.67	2.42
3	Holt-Winters	30	0.012911	38.78	2.49
4	ARIMA	42	0.018233	86.4	2.54
5	ARIMA	38	0.017507	63.93	2.7
6	Holt-Winters	32	0.045307	90.75	2.78
7	Holt-Winters	44	0.011293	64.93	2.8
8	SARIMA	36	0.011081	64.94	2.85
9	Holt-Winters	4	0.085294	90.1	2.98
10	Holt-Winters	26	0.042117	86.7	3.03
11	Holt-Winters	34	0.039903	86.84	3.25
12	Holt-Winters	8	0.039805	85.88	3.3
13	Holt-Winters	33	0.011109	67.93	3.32
14	Holt-Winters	13	0.100067	87.34	3.38
15	Holt-Winters	41	0.062456	88.81	3.42

Select Store: 37 - MAPE: 2.24 **Model:** Holt-Winters(0.12, 0.02, 0.77) **MAPE:** 2.24% **RMSE:** $15,054 **R-Squared:** 55.97%





Model Results

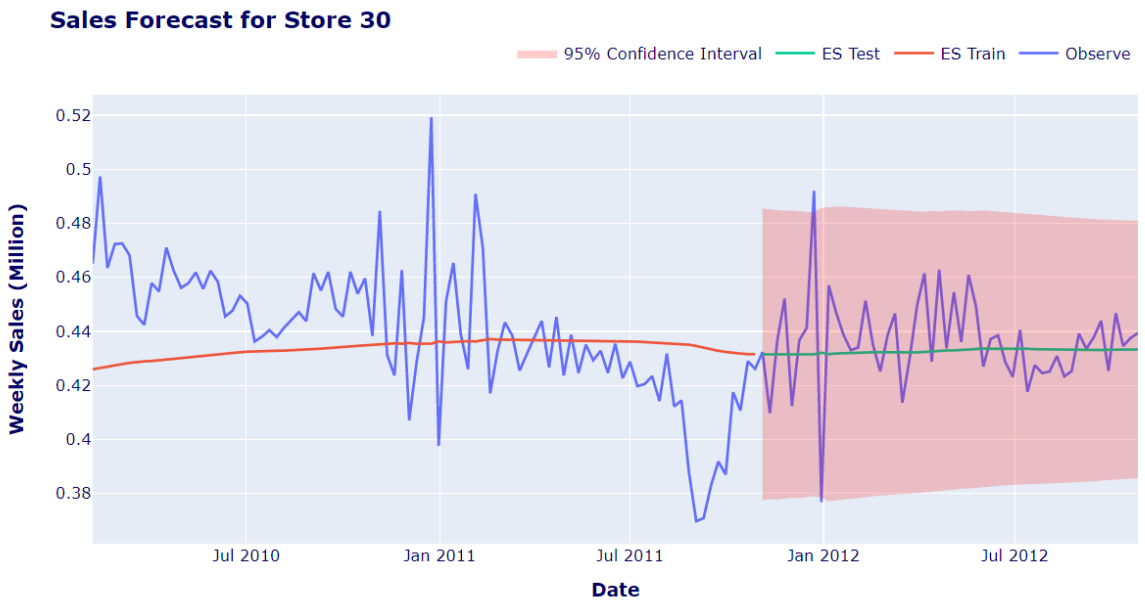
Holt-Winters outperform other models in making sales forecast for most stores.

Exponential Smoothing and ***Holt's Linear Trend*** models are not selected in making sales forecast for all stores.

Model	Number of Models Selected in Making Sales Forecast	RMSE Range	MAPE Range	R-Squared Range
Holt-Winters	37	***$11,109*** to ***$160,250***	***2.23%*** to ***8.01%***	***-56.16%*** to ***90.75%***
ARIMA + SVD	4	***$11,296*** to ***$360,007***	***2.3%*** to ***40.15%***	***-290.13%*** to ***86.4%***
SARIMA + SVD	4	***$11,081*** to ***$311,899***	***2.36%*** to ***16.53%***	***-27.13%*** to ***87.36%***
Exponential Smoothing	0	***$15,241*** to ***$305,614***	***2.73%*** to ***10.61%***	***-13.14%*** to ***33.67%***
Holt's Linear Trend	0	***$14,813*** to ***$305,646***	***2.99%*** to ***10.27%***	***-28.62%*** to ***37.35%***

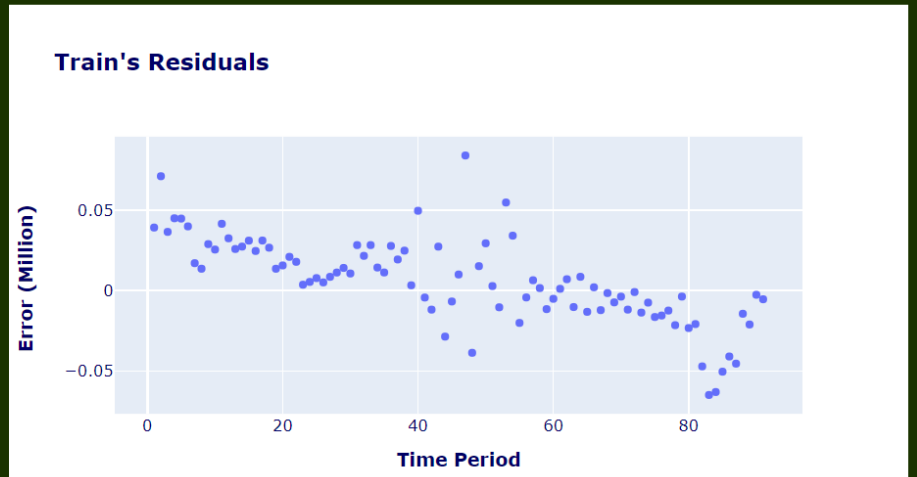
Simple Exponential Smoothing Model

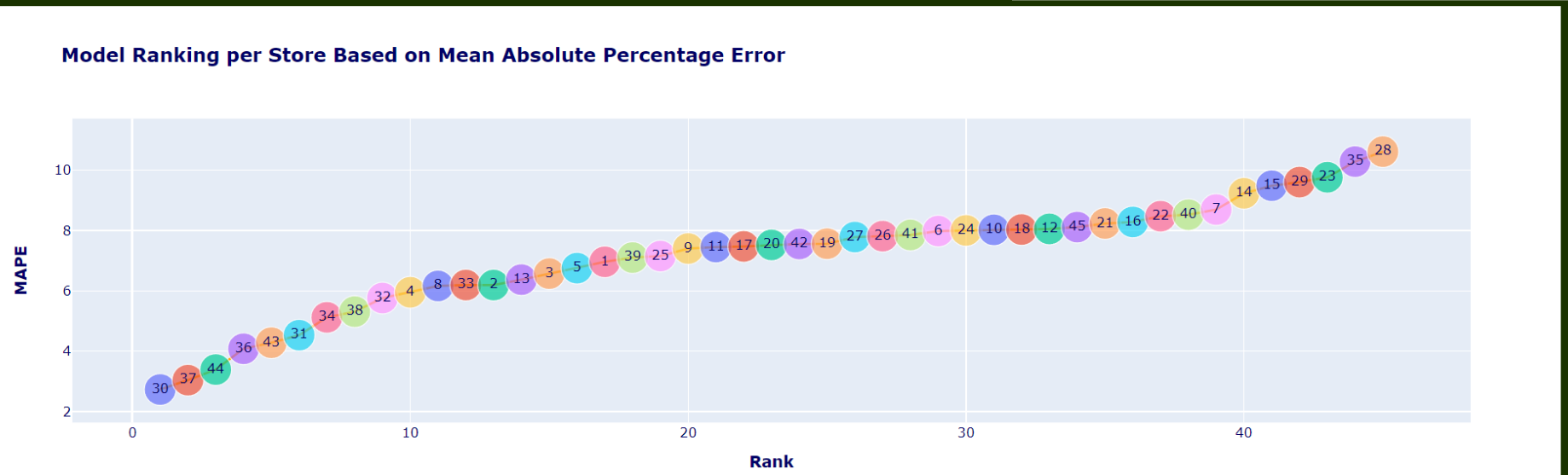
Select Store: 30 - MAPE: 2.73 **Alpha:** 0.01 **Rank model by:** MAPE RMSE R-Squared **Page Load Status:** Completed



Evaluation Metrics

Train Data	MAPE:	4.84%	R-Squared:	-15.01%	RMSE: $27,373
Test Data	MAPE:	2.73%	R-Squared:	-6.16%	RMSE: $17,001



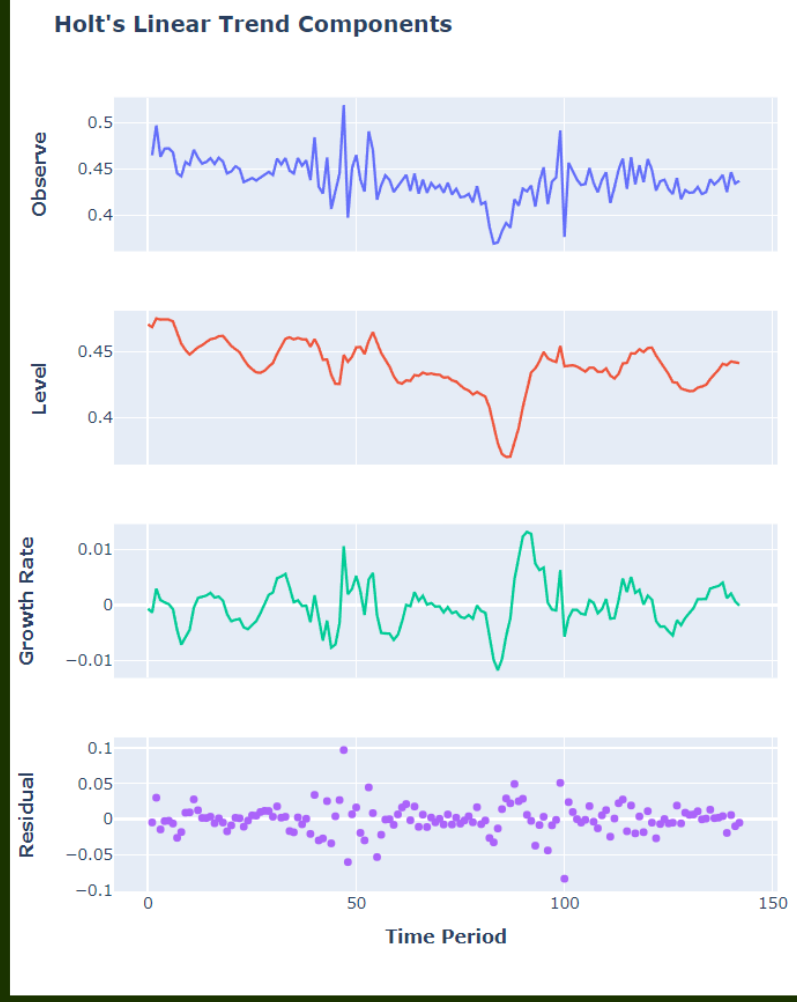


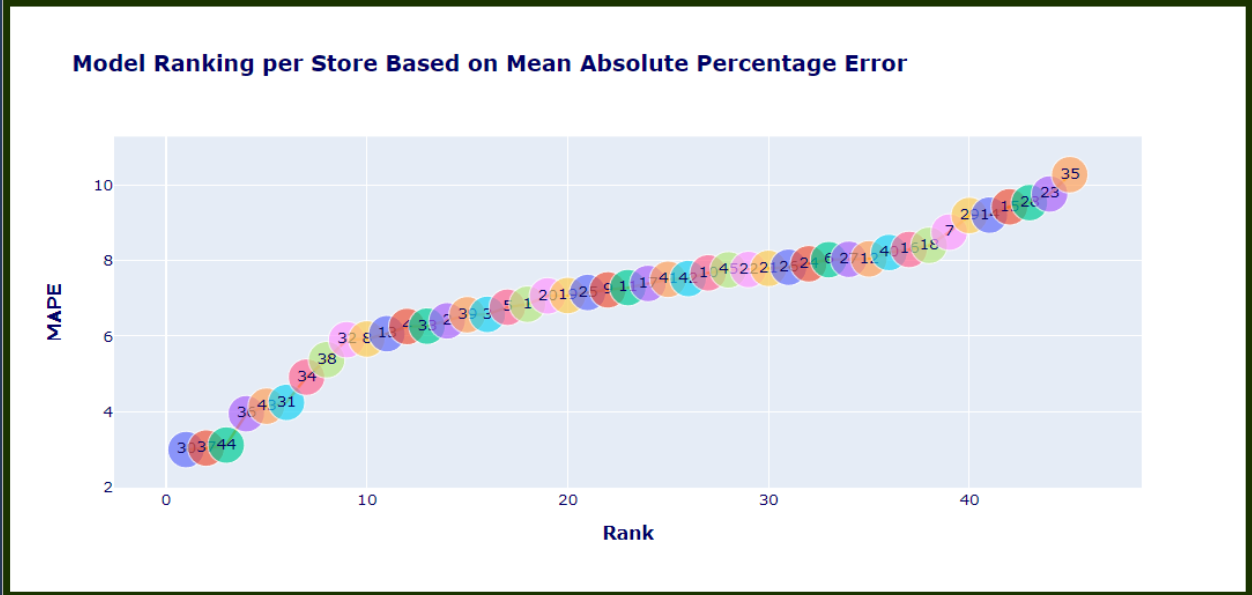
Holt's Linear Trend Model

Select Store: 30 - MAPE: 2.99 **Alpha:** 0.26 **Gamma:** 0.55 **Rank model by:** (•) MAPE ( ) RMSE ( ) R-Squared **Page Load Status:** Completed

Evaluation Metrics

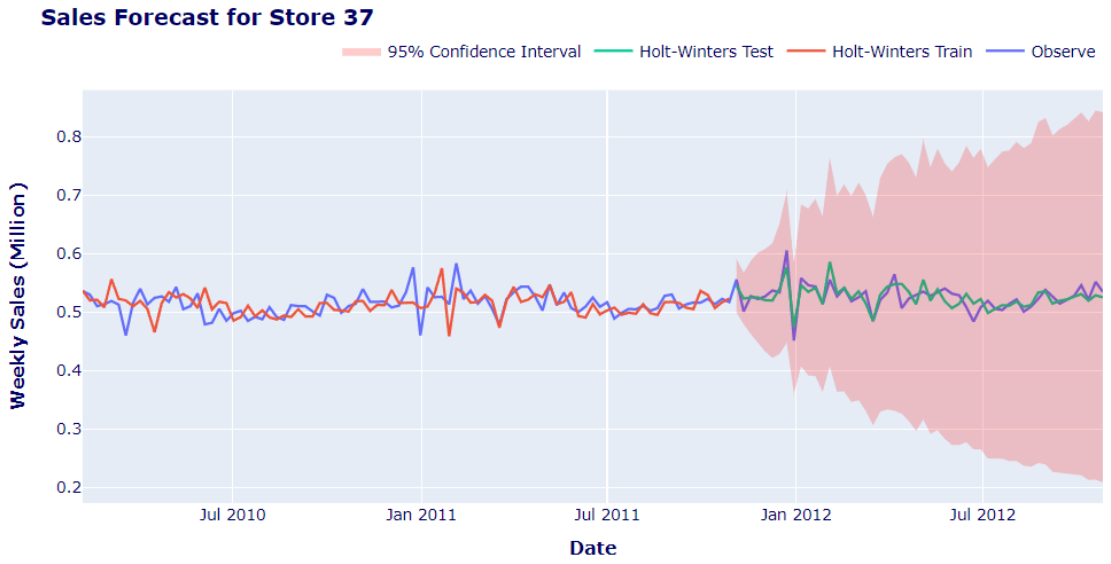
Train Data	MAPE:	3.32%	R-Squared:	31.51%	RMSE: $21,122
Test Data	MAPE:	2.99%	R-Squared:	-42.84%	RMSE: $19,721

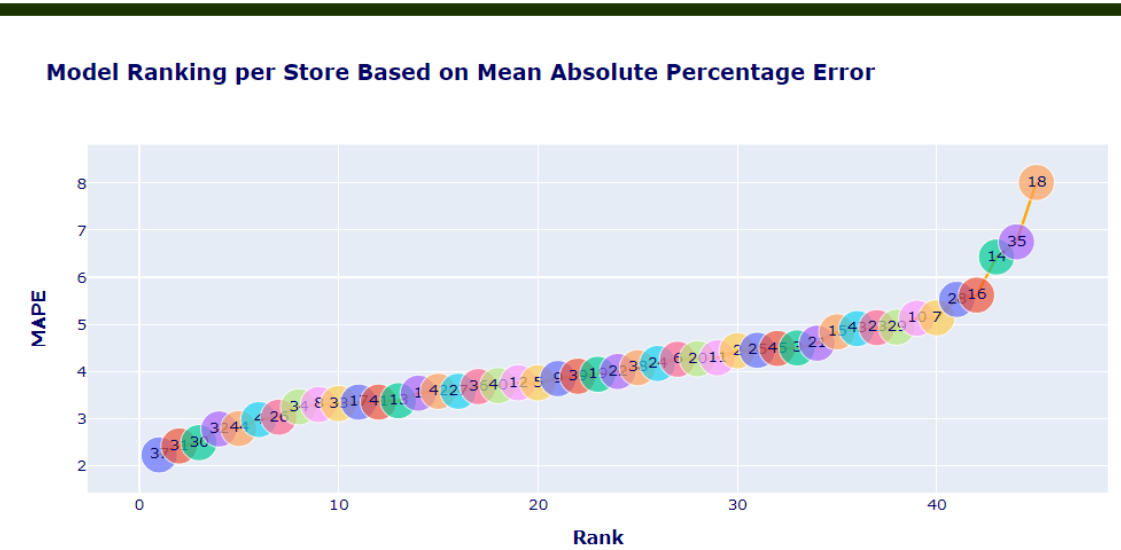




Multiplicative Holt-Winters Model

Select Store: 37 - MAPE: 2.23 **Alpha:** 0.01 **Gamma:** 0.03 **Delta:** 0.86 **Rank model by:** (•) MAPE ( ) RMSE ( ) R-Squared **Page Load Status:** Completed



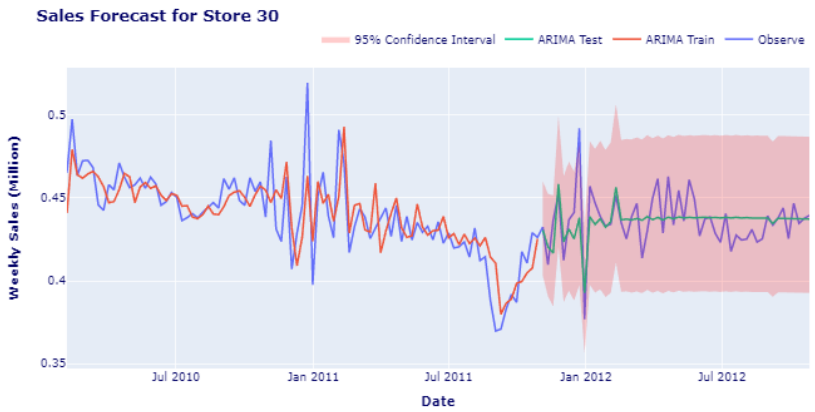


Evaluation Metrics

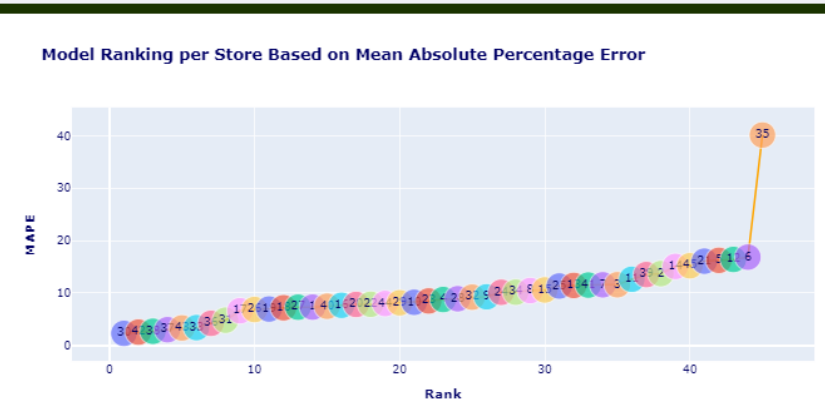
Train Data	MAPE: 3.17%	R-Squared: -18.42%	RMSE: $21,732
Test Data	MAPE: 2.23%	R-Squared: 55.58%	RMSE: $15,122

(Seasonal) Autoregressive Integrated Moving Average Plus SVD Model

Model: ◉ ARIMA ○ SARIMA Select Store: 30 - MAPE: 2.3 p: 12 d: 0 q: 7 P: 0 Q: 0 SVD: 4 Rank model by: ○ AIC ◉ MAPE ○ RMSE ○ R-Squared

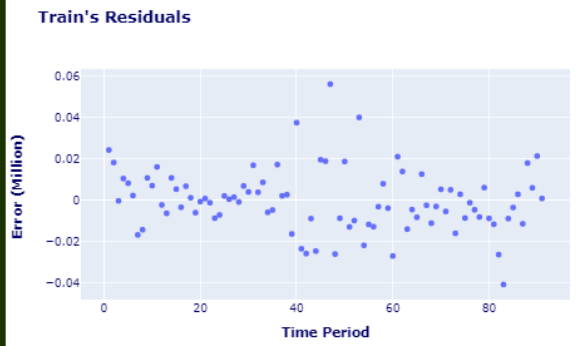


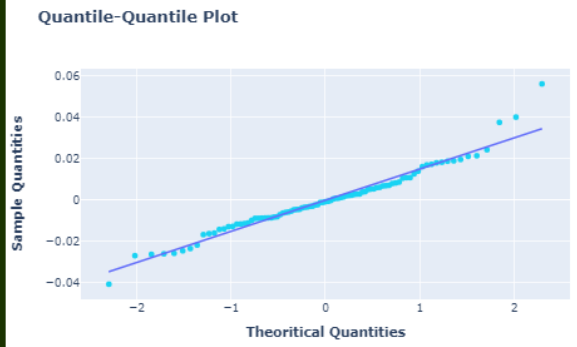

```
==============================================================================
Dep. Variable:     Log of Weekly Sales (Million)   No. Observations:                   91
Model:                      ARIMA(12, 0, 7)   Log Likelihood                 178.172
Date:                      Wed, 19 Apr 2023   AIC                           -306.343
Time:                              22:39:39   BIC                           -243.572
Sample:                                   0   HQIC                          -281.019
                                        - 91
Covariance Type:                        opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.8181      0.037    -22.207      0.000      -0.890      -0.746
x1            -0.0104      0.047     -0.221      0.825      -0.102       0.082
x2             0.0476      0.029      1.639      0.101      -0.009       0.104
x3             0.0013      0.037      0.036      0.971      -0.070       0.073
x4            -0.1110      0.035     -3.152      0.002      -0.180      -0.042
ar.L1          0.1530      3.736      0.041      0.967      -7.170       7.476
ar.L2          0.1201      2.736      0.044      0.965      -5.243       5.483
ar.L3          0.0235      1.756      0.013      0.989      -3.419       3.466
ar.L4          0.1118      1.379      0.081      0.935      -2.591       2.814
ar.L5         -0.1533      1.848     -0.083      0.934      -3.776       3.469
ar.L6          0.2232      1.241      0.180      0.857      -2.208       2.655
ar.L7          0.0008      1.454      0.001      1.000      -2.849       2.850
```

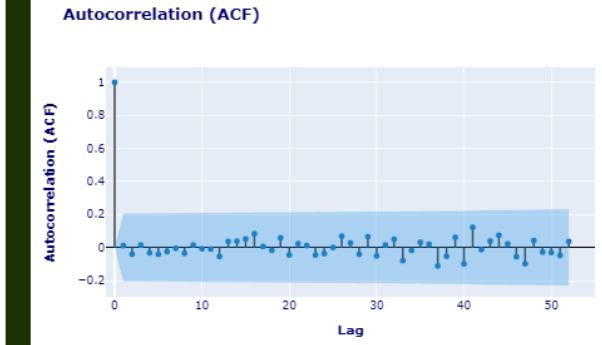


Evaluation Metrics

Train Data	MAPE: 2.54%	R-Squared: 65.1%	RMSE: $15,079
Test Data	MAPE: 2.3%	R-Squared: 30.9%	RMSE: $13,717







Conclusion

- Selected forecast models have ***less than 10% MAPE***.
- ***Holt-Winters*** models
  - Less complex than ARIMA and SARIMA models in most cases
  - Outperform other models in most cases
- ***Exponential Smoothing*** and ***Holt's Linear Trend*** models are
  - Simple but do not perform well in most cases
  - Sales forecast is not reliable beyond one period into the future
- Sales forecast application could be extended by
  - Including machine learning and deep learning models such Random Forest, Recurrent Neural Networks, etc.
  - Forecasting sales for other retail stores, not limited to Walmart stores.

References

1. Trent, C. (2022, February 7). *9 Eye-Opening Sales Forecasting Statistics You Need to Know in 2022*. Dooly. https://www.dooly.ai/blog/sales-forecasting-statistics/
2. *Don't Ignore these Five Sales Forecasting Stats*. (2021, July 23). InsightSquared. https://www.insightsquared.com/blog/dont-ignore-five-forecasting-stats/
3. Bevans, R. (2020, March 26). Akaike Information Criterion | When & How to Use It. Scribbr. https://www.scribbr.com/statistics/akaike-information-criterion/
4. Mendenhall, W. (2019). *SECOND COURSE IN STATISTICS: regression analysis*. Prentice Hall.
5. Bowerman, B. L., O'connell, R. T., & Koehler, A. B. (2005). *Forecasting, time series, and regression: an applied approach*. Thomson Brooks/Cole.
6. Peixeiro, M. (2022). *Time Series Forecasting in Python*. Simon and Schuster.
7. Hyndman, R., Koehler, A. B., J Keith Ord, Snyder, R. D., & Springerlink (Online Service. (2008). *Forecasting with Exponential Smoothing: The State Space Approach*. Springer Berlin Heidelberg.
8. Verma, Y. (2021, August 18). *Complete Guide To Dickey-Fuller Test In Time-Series Analysis*. Analytics India Magazine. https://analyticsindiamag.com/complete-guide-to-dickey-fuller-test-in-time-series-analysis/