TimeSeries Walmart Project

Kenya Roy and Lani Lewis

2024-12-02

# EDA & CLEAN DATA

## Merge Data

### Validation Dataset

### Clean up merge duplicate columns

### Clean up date types

**Training Dataset**

**Validation Dataset**

## NA Values

## Store Date Dept Weekly\_Sales IsHoliday Temperature   
## 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000   
## Fuel\_Price MarkDown1 MarkDown2 MarkDown3 MarkDown4 MarkDown5   
## 0.00000 64.25718 73.61103 67.48085 67.98468 64.07904   
## CPI Unemployment Type Size   
## 0.00000 0.00000 0.00000 0.00000

### Remove NA >50%

### SPLIT

**Filtered Training Dataset**

## Aggregate Weekly\_Sales and retain other columns

## 80/20 Train Test Split

## Training set dimensions: 114 8

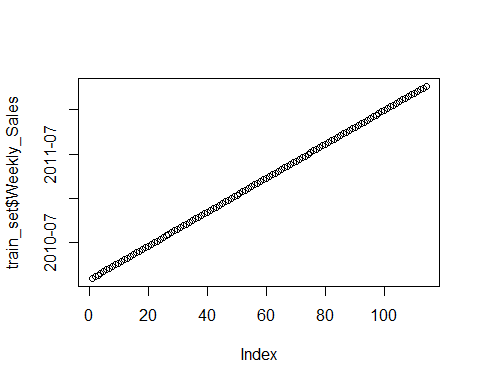
## Test set dimensions: 29 8

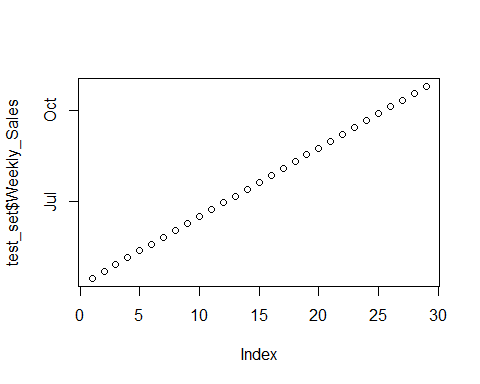
## Training set percentage: 79.72%

## Test set percentage: 20.28%

#### TEST SPLIT OF DATA

**Aggregated Date**

**Train Test Set** 

**Test Set** - This is different then the validation set 

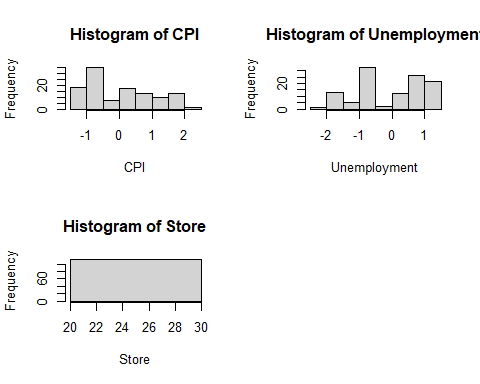
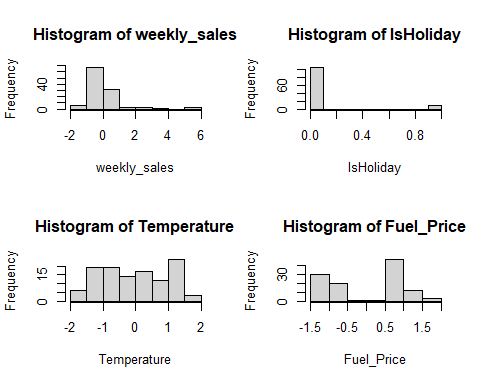
#### SCALE VARIABLES

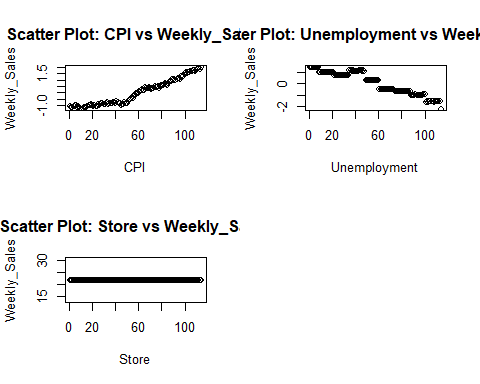
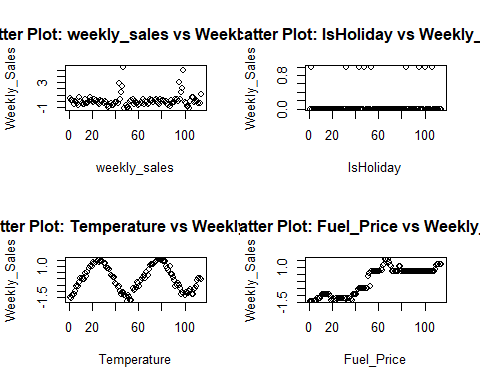
**TRAINED**

**TEST**

##### COVERT DATASETS TS

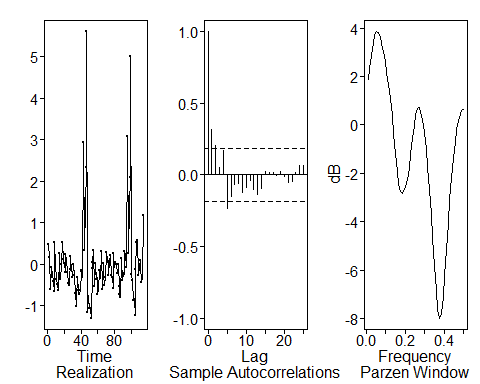
## PLOTS

**Variables that are highly skewed often benefit from a log transformation.** 

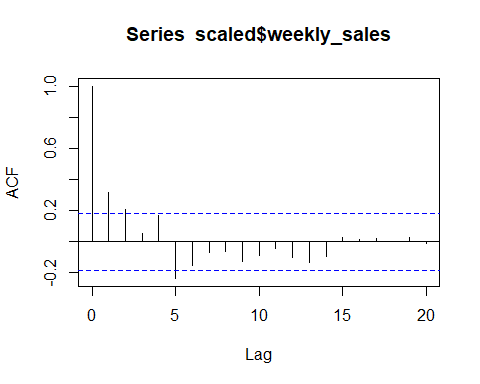
**Look for non-linear relationships where a log transformation could help linearize the data.** - Possibly log variables 

**May not need to use Highly correlated data** - Keep all variables

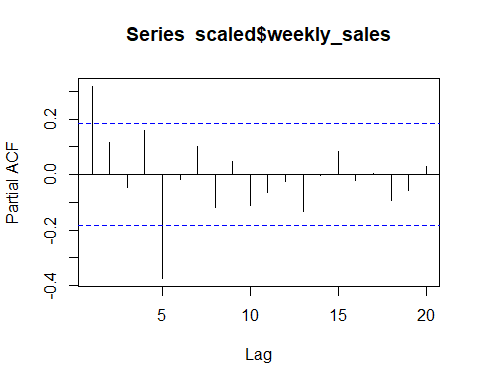
### TS Pots

* Realization: The sharp spikes and upward or downward trends suggest that the series is non-stationary (mean and variance are not constant over time).
* ACF: Strong correlations at lag 1 and lag 2, gradually decreasing, also indicating non-stationarity.
* Spectral Density: looks like we have three high peaks the first is the strongest and we see high peaks around period 12 (13/14)
* The next import periods are around 3 then 2…
* The overall behavior indicates that differencing might be required to stabilize the series. 

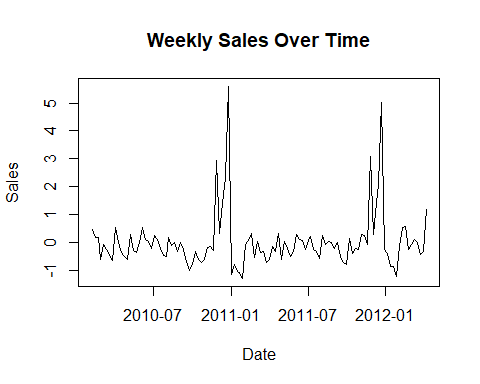
#### ACF | MA(q) Review

* Maybe MA(4) 

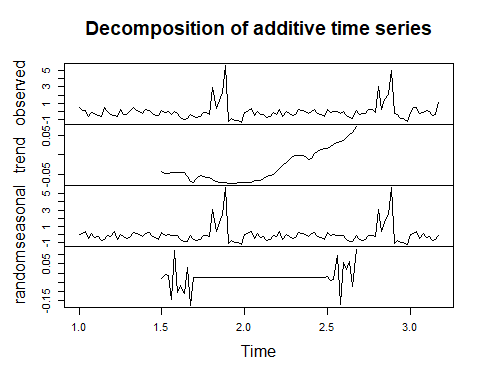
#### PACF | AR(p) Review

* maybe a AR(4)
* scaled it looks like an AR(2) 

#### Plot for Trend

* It looks like there are spike around 13(11 - Nov) and 14(12 - Dec) period?
* Maybe a frequency peak at around .071 - .077 

##### Decompose Time Series

**Summary of Insights** - **Trend**: A clear upward trend is present between time 2 and 3, indicating long-term growth. - **Seasonality**: The data has a strong seasonal component, which should be explicitly modeled. - **Stationarity**: The presence of both trend and seasonality suggests the series is non-stationary. First-order differencing (d=1) and/or seasonal differencing (D=1) may be required to stabilize the series for ARIMA modeling. - **Outliers**: The sharp spikes in the residuals suggest potential outliers or irregular events that may require further investigation or adjustments. — 

#### Dickey-Fuller Test for Stationarity

* Reject the null hypothesis of Non-Stationarity
* This test was not as helpful as I had hopped cause we know that we need to difference the model from the plots above.
* This is actually a non-stationary model based off Kaggle as well. On the site they already mention this is a non-stationary model

# ARMA MODEL

## AIC Model Determination

* I will test the following three models
  + ARMA(5,1) increase the values for more options since we are at the max
    - ARMA(5,1) is still in the top 5 so I will try this option
    - ARMA(5,3) and ARMA(6,1) will be the next I play with.
* After testing I found that ARMA(5,3) performs the best

## BIC Model Determination

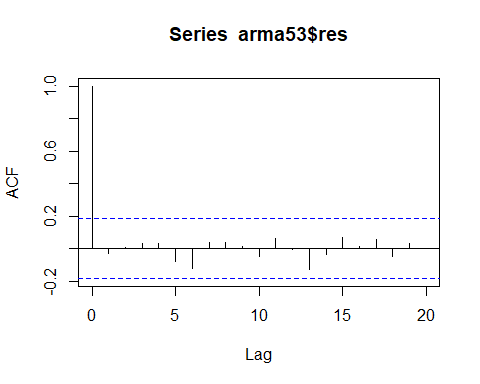
* Not going to use this model cause we know this is not white noise. So I increased the model
* Still not valid options

## ARMA(5,3) Factor Table Check | Model Estimates

* Here the 1-B is very apparent. This might be the best model to use since I know I need to difference my model
* The frequency also seems to match up almost exactly with what I am seeing in the original realization

### RESIDUAL CHECKS ARMA(5,3)

#### ACF Residual Check

* Passed the ACF White Noise Residual Check 

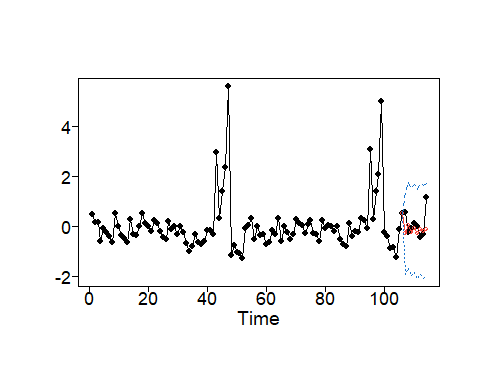
#### Ljung Residual Check

* Fail to reject the Null Hypothesis of White Noise as the p-value is greater then .05 for both tests
* This is another pass for this model

## p-value 0.8608067

## p-value 0.9770861

## Forecast



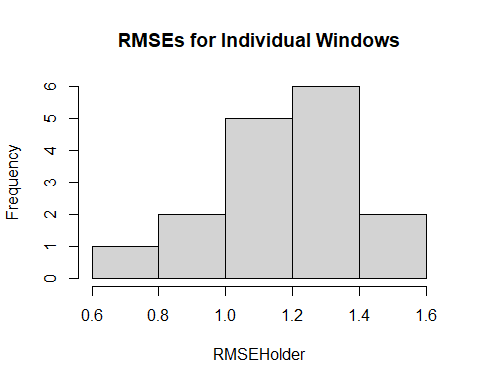
### ASE

## [1] 1.014748

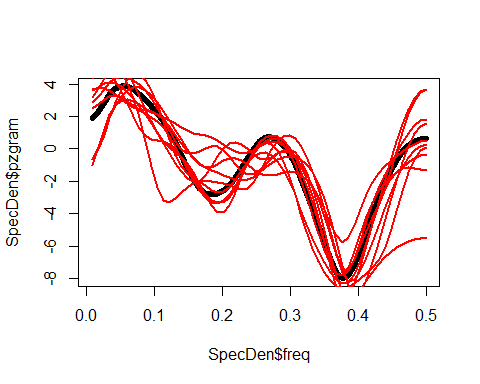
### WMAE | Kaggle

## [1] 0.8192932

### Rolling Window RMSE



### Compare Multiple Spectral Densities

* This model appears to perform well with generating the spectral densities 

### Compare Multiple ACFs

* Does fairly well modeling the ACFs 