Retail Sales Analysis

A project on USA Retail sales Analysis

Table of Content

Project objective

Background and context

USA Retail sales data

Data Due Diligence

Analysis of Data

Modeling The Time Series

Project objective

Forecasting Sales Data Using Statistical Modeling and Machine Learning

Here we are using US sales data across various categories from 1992 - 2020, to predict future sales and find out the effect of some crisis along the way

Background and context

USA Retail sales data

1. The data was collected from the US census Bureau. Census Bureau monitored response and data quality and determined estimates for various sector of USA economy.
2. The data collected is from 1992 - 2020. Here we are using Time-series Analysis to find insights and forecast the sales estimates.
3. We also find the impact of various major crisis along the way such as 2008 Economic crisis, and more recently COVID-19 and it impact on retail sector
4. We have done various tasks for this data such as data collection, cleaning, visualization, data analysis, Prediction (both statistical methods and Machine learning models for forecasting)

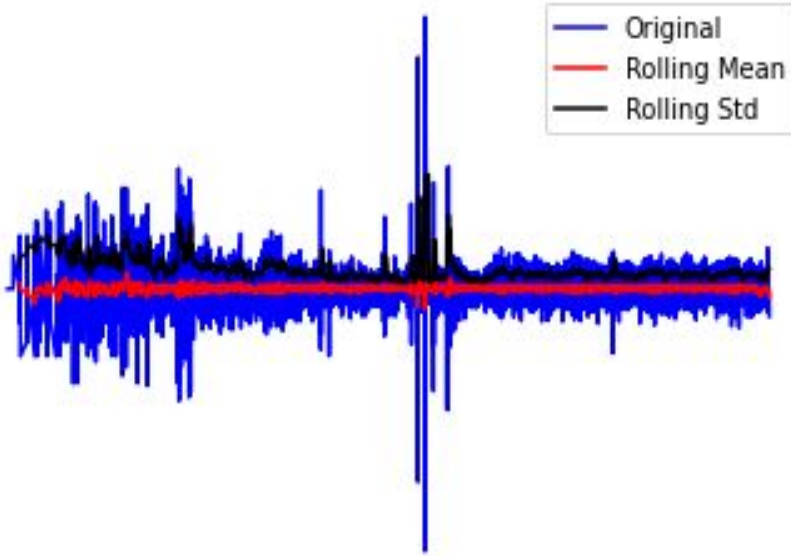
Data Due Diligence

Time-Series Analysis

Time-series data can be defined in many ways. A simple definition can be "data collected on the same metric or same object at regular or irregular time intervals."

Do we need Time series Analysis:

We should think about whether there is an inherent relationship or structure between data at various time points (e.g., is there a time-dependence), and whether we can leverage that time-ordered information.



Understanding the data

The dataset contains the estimates of Monthly Retail and Food Services Sales by Kind of Business from the year 1992 - 2020. These estimates are shown in millions of dollars and are based on data from the Monthly Retail Trade Survey, Annual Retail Trade Survey, * Service Annual Survey, and administrative records.

The dataset contains both adjusted and unadjusted for seasonal variations for various categories. These categories shows various kind of Business categories operating in USA. These categories are based on North American Industry Classification System (NAICS).

Data Levels

- The data was organised using North American Industry Classification System (NAICS) coding system. It contains multi level categorization of Businesses including:
  - **Top level Sales Estimates:** Total Estimates of Sales of US as a whole. Ex. 44X72: Retail Trade and Food Services: U.S. Total
  - **First Level :** Sales Estimates of categories. Ex. 441:Motor vehicle and parts dealers, 442:Furniture and home furnishings stores
  - **Second level :** Ex. 4411: Automobile dealers, 4413: Automotive parts, acc., and tire stores.
  - **Third Level:** Ex. 44111: New car dealers, 44112: Used car dealers
- We need to separate these categories and perform Time series Analysis on each particular levels separately

Handling inconsistencies in our data

- Following are the inconsistencies that was in our data:
  - (S) Suppressed - Estimate does not meet publication standards because of high sampling variability (coefficient of variation is greater than 30%), poor response quality (total quantity response rate is less than 50%), or other concerns about the estimate quality.
  - (NA) Not available
- Since our data is Estimate of sales we have suppressed values in our data. We can handle these values by predicting these values from the past values
- Not available values can also be predicted using the same techniques we used for Suppressed values
- Handling the different data types and converting them into datetime and float64 was also one of the tasks

What questions do we want to Answer?

- I think that it is important to start with a very concrete question that we think can be answered with data, specifically with time-series data.
- Some of the questions that we might answer with this forecasting:
  - Forecast future sales Estimate of retail sector of USA.
  - Forecast future sales Estimate of different categories in retail Sector.
- Understand the limitations of the data and what potential questions can be answered by data is important. These questions can reduce, expand, or modify the scope of our project.

What techniques may help us to answer these Questions ?

Statistical models

- Ignore the time-series aspect completely and model using **traditional statistical modeling toolbox**. Examples. Regression-based models.
- **Univariate statistical time-series modeling.** Examples. Averaging and smoothing models, ARIMA models.
- **Slight modifications to univariate statistical time-series modeling.** Examples. External regressors, multivariate models.
- **Additive or component models**. Examples. Facebook Prophet package.
- **Structural time series modeling.**Examples. Bayesian structural time series modeling, hierarchical time series modeling.

What techniques may help us to answer these Questions ?

Machine learning models

- Ignore the time-series aspect completely and model using **traditional machine learning modeling toolbox.** Examples. Support Vector Machines (SVMs), Random Forest Regression, Gradient-Boosted Decision Trees (GBDTs), Neural Networks (NNs)
- **Hidden markov models (HMMs).**
- **Other sequence-based models.**
- **Gaussian processes (GPs).**
- **Recurrent neural networks (RNNs).**

What techniques may help us to answer these Questions ?

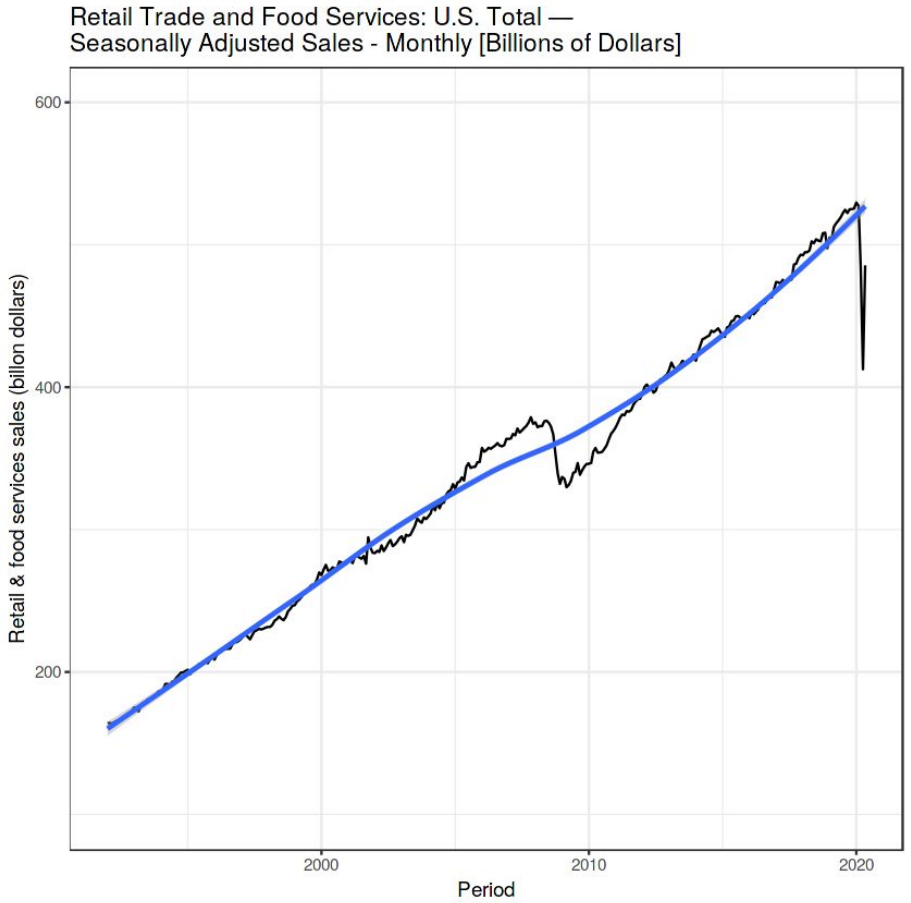
Additional data considerations before choosing a model

- Whether or not to incorporate external data
- Whether or not to keep as univariate or multivariate (i.e., which features and number of features)
- Outlier detection and removal
- Missing value imputation

Analysis of Data

Plotting the data

- There does appear to be an overall increasing trend.
- There appears to be some differences in the variance over time.
- There is seasonality (i.e., cycles) in the data.
- There are outliers.



Look at Stationarity

- Most of the time-series model that we use assume the stationarity of time-series. This assumption gives us some nice statistical properties thereby allowing us to use various models for forecasting.
- To put it in layman terms, if we want to predict future using the past data, w should assume that the data will follow the same trends and patterns as in the past. This general statement holds for most training data and modeling tasks.

Look at Stationarity

Stationary Time-Series have following characteristics:

- Constant mean
- Constant variance
- Autocovariance doesn't depend on time

Sometimes in order to make time-series stationary we need to transform the data. However, this transformation then calls into questions if this data is truly stationary and is suited to be modeled using these techniques.

Dickey-Fuller Test

We can test Stationarity using moving average statistics and Dickey-Fuller Test. Following are the conditions for Hypothesis testing using Dickey Fuller Test

- Null Hypothesis (H_0): time series is not stationary
- Alternative Hypothesis (H_1): time series is stationary


```
Dickey-Fuller Test:
Null Hypothesis (H_0): time series is not stationary
ALternative Hypothesis (H_1): time series is stationary
Results of Dickey-Fuller Test:
Reject the null hypothesis (H0), the data does not have a unit root and is stationary.
Test Statistics                -1.405752e+01
p-value                         3.098177e-26
# Lags Used                    3.900000e+01
Number of observation Used     1.019200e+04
Critical Value (1%)           -3.430992e+00
Critical Value (5%)           -2.861824e+00
Critical Value (10%)          -2.566921e+00
dtype: float64
```

Handling Stationarity

It is common for a time-series to have Non stationary behaviour. Most common reason behind non- stationary time-series are:

- Trend - mean is not constant over time .
- Seasonality - variance is not constant over time

There are ways to correct for trend and seasonality in order to make times-series stationary.

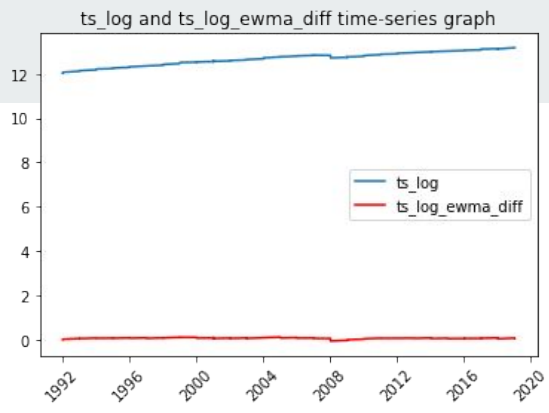
What will happen if we don't correct stationarity?

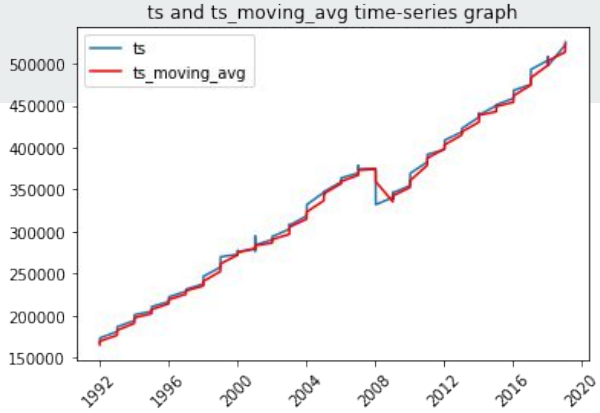
Many things can happen:

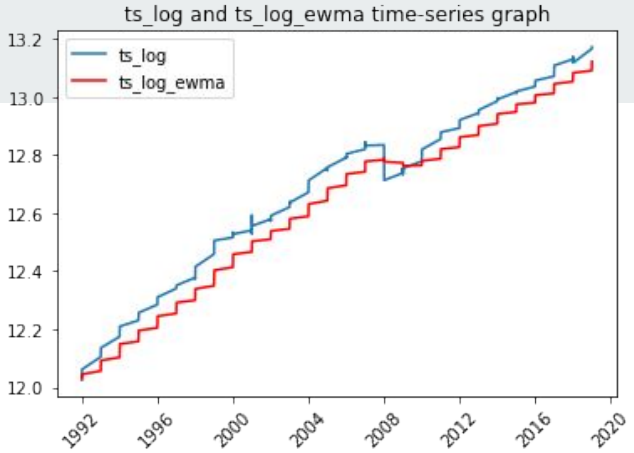
- Variance can be mis-specified
- Model fit can be worse
- Not leveraging valuable time-dependent nature of data.

Eliminating Trend And Seasonality

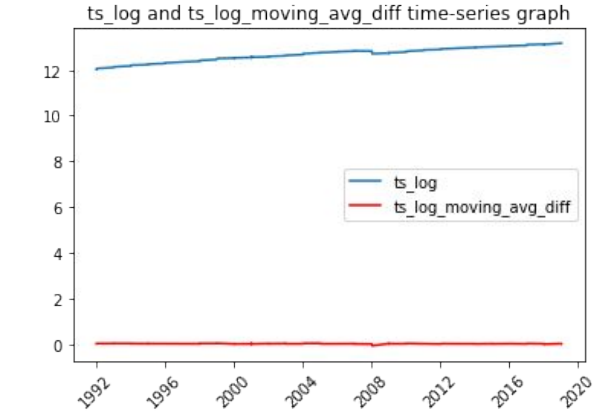
- **Transformation**
  - Examples. Log, square root, etc.
- **Smoothing**
  - Examples. Weekly average, monthly average, rolling averages.
- **Differencing**
  - Examples. First-order differencing.
- **Polynomial Fitting**
  - Examples. Fit a regression model.
- **Decomposition**

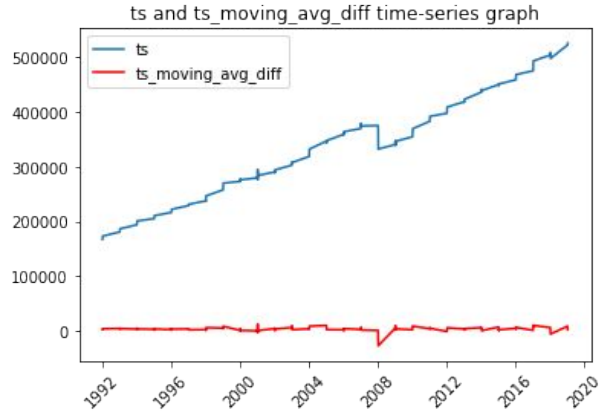


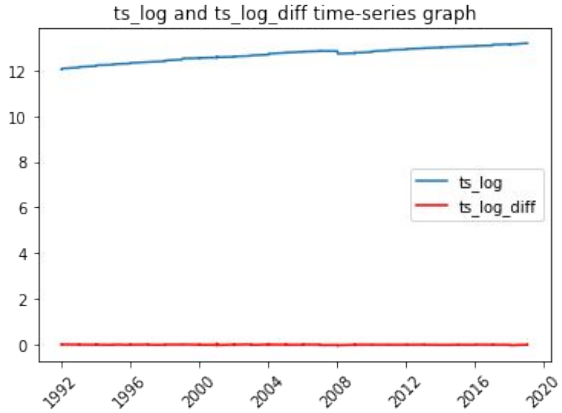




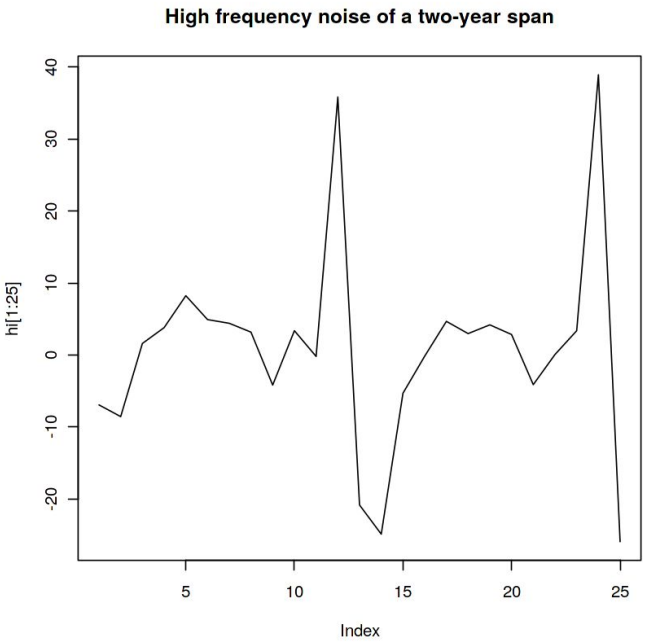
Transformation, Smoothing and Differencing of Time Series

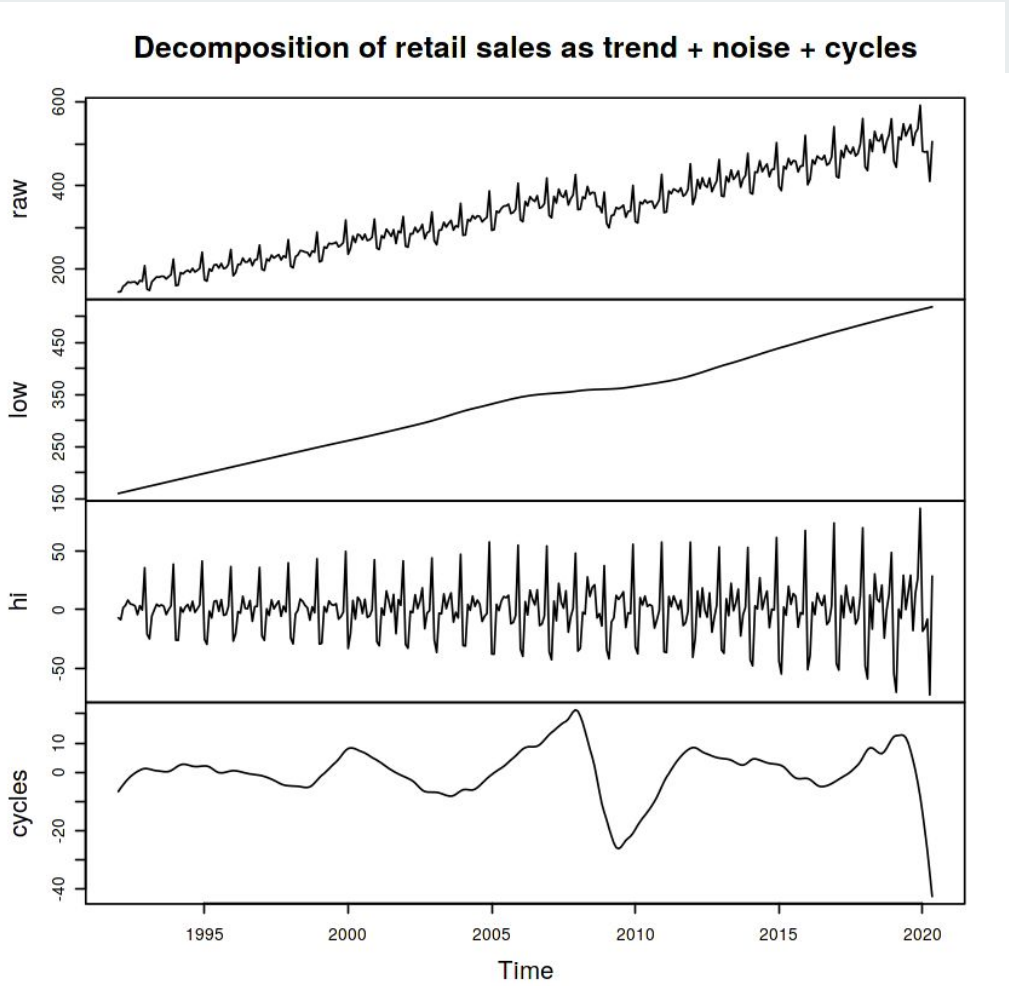






Decomposition of Time Series





For 44X72: Retail Trade and Food Services: U.S. Total

- The low frequency plot shows us an estimate of the trend followed by the sales from 1992 to 2020. The average increase rate is 12.795 billion dollars per year
- The high frequency plot shows us the seasonal changes in Retail sales. When I zoom into high frequency we see a sales peak in December and also at around May and June and this pattern occurs over every year. The seasonal change have a period of one year
- The middle frequency plot tells us about any long term changes is their is any. I can see that there is a steady increase from 2004 to 2008, but from around 2nd quarter of 2008 retail sales begin to decline and it is not until last quarter of 2009 it began to climb back. This shows the stock market crash on Sept. 29, 2008.The decline in Retail Sales partly reflects the economic downturn.The same is happening in the 1st quarter of 2020 due to COVID-19 outbreak which is still going on. The decline in Retail Sales shows people refrain from spending and shows the lockdown situation of the country

Modeling The Time Series

Why is statistical forecasting important (or at least, interesting)?

"Forecasting can take many forms—staring into crystal balls or bowls of tea leaves, combining the opinions of experts, brainstorming, scenario generation, what-if analysis, Monte Carlo simulation, solving equations that are dictated by physical laws or economic theories—but statistical forecasting, which is the main topic to be discussed here, is the art and science of forecasting from data, with or without knowing in advance what equation you should use."

Robert Nau, Principles and Risks of Forecasting

ARIMA Model (Autoregressive Moving Average)

We can use Arima model when we know there is dependence between values and we can leverage that information to forecast.

Assumptions = The time-series is stationary

ARIMA depends on following 3 terms:

1. Number of AR (Auto-Regressive) terms (p).
2. Number of I (Integrated or Difference) terms (d).
3. Number of MA (Moving Average) terms (q).

ARIMA Model (Autoregressive Moving Average)

How do we determine p, d, and q? For this, we can use ACF and PACF plots

- **Autocorrelation Function (ACF),** correlation between the time series with a lagged version of itself(Ex: Correlation of Y(t) with Y(t-1)).
- **Partial Autocorrelation Function (PACF)**, Additional correlation explained by each successive lagged terms.

How to interpret ACF and PACF plots?

- p - Lag value where the PACF chart crosses the upper confidence interval for the first time.
- q - Lag value where the ACF chart crosses the upper confidence interval for the first time.

ACF and PACF Functions:

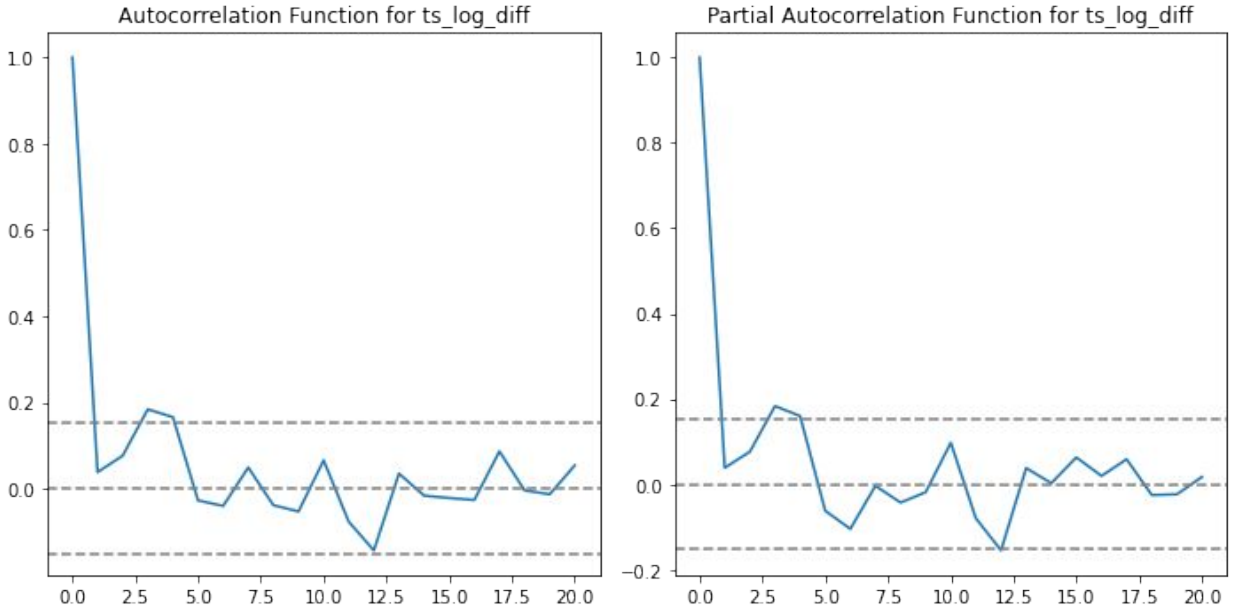
How do we determine p, d, and q? For this, we can use ACF and PACF plots

- **Autocorrelation Function (ACF),** correlation between the time series with a lagged version of itself(Ex: Correlation of Y(t) with Y(t-1)).
- **Partial Autocorrelation Function (PACF)**, Additional correlation explained by each successive lagged terms.

How to interpret ACF and PACF plots?

- p - Lag value where the PACF chart crosses the upper confidence interval for the first time.
- q - Lag value where the ACF chart crosses the upper confidence interval for the first time.

For 44X72: Retail Trade and Food Services: U.S. Total



SARMA(Seasonal Autoregressive moving Average

- When we see a season in out time series we can use SARMA
- We can see a seasonal period of 12 on the retail sales data. A SARMA model with period equal to 12 can be used.
- We use SARMA model on log transformed Data.
    - p = 1, q = 1 i.e. SARMA(p,q) $\times(1,1)12$ model was the best for total sales

Output

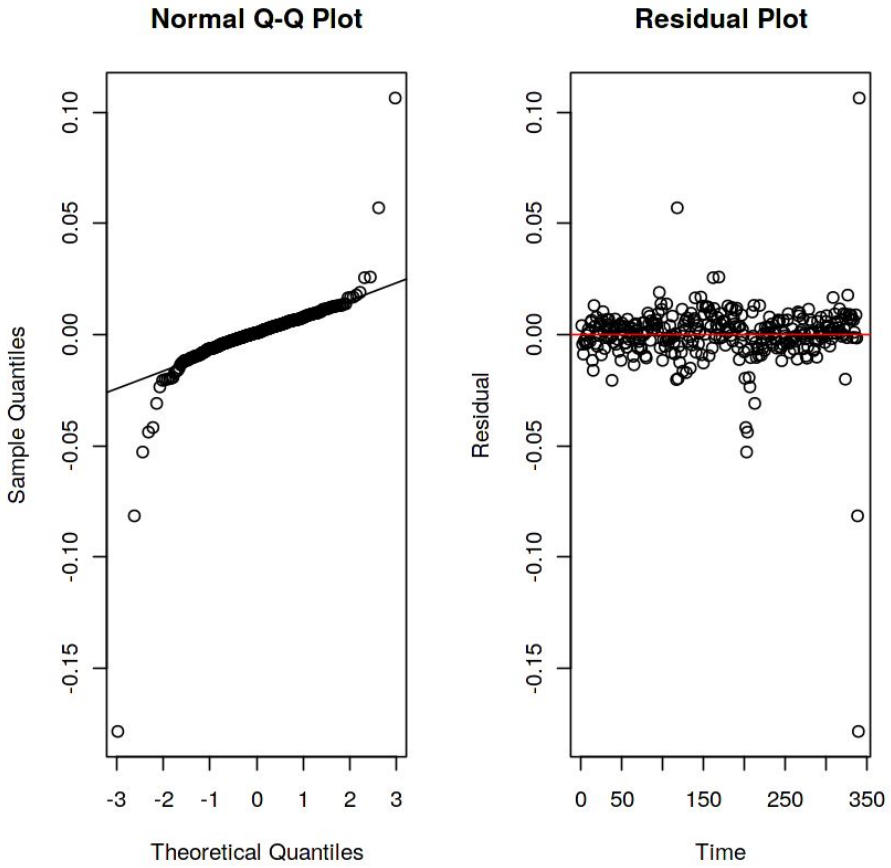
Call:

arima(x = new_log$detrend, order = c(2, 0, 2), seasonal = list(order = c(1, 0, 1), period = 12))

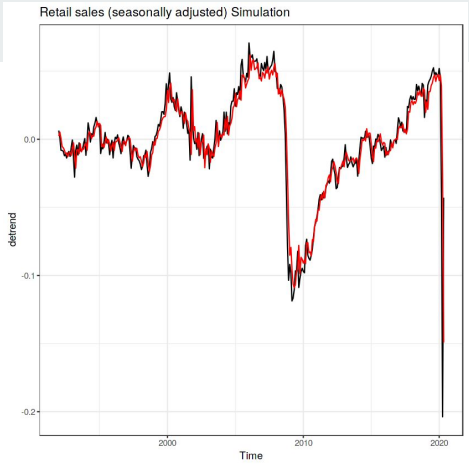
Coefficients:

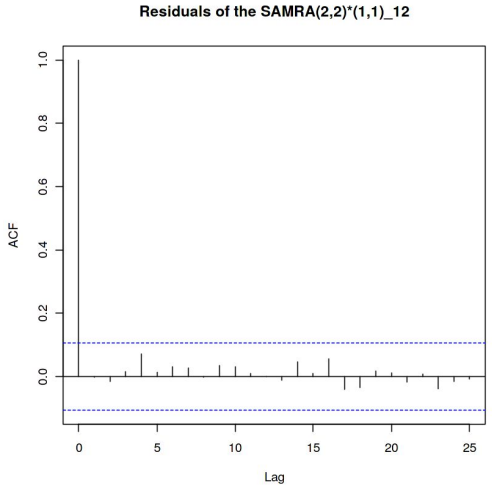
	ar1	ar2	ma1	ma2	sar1	sma1	intercept
	0.7573	0.2006	0.0130	-0.2178	0.7016	-0.8625	-0.0029
s.e.	0.2892	0.2768	0.2754	0.0885	0.1604	0.1177	0.0077

sigma^2 estimated as 0.0002321: log likelihood = 941.04, aic = -1866.09


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-3
-2
-1
0
1
2
3
0.10
0.05
0.00
-0.05
-0.10
-0.15

Residual Plot
Residual
Time
0
50
150
250
350
0.10
0.05
0.00
-0.05
-0.10
-0.15


Retail sales (seasonally adjusted) Simulation
detrend
Time
0.0
-0.1
-0.2
2000
2010
2020


Residuals of the SAMRA(2,2)*(1,1)_12
ACF
Lag
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25

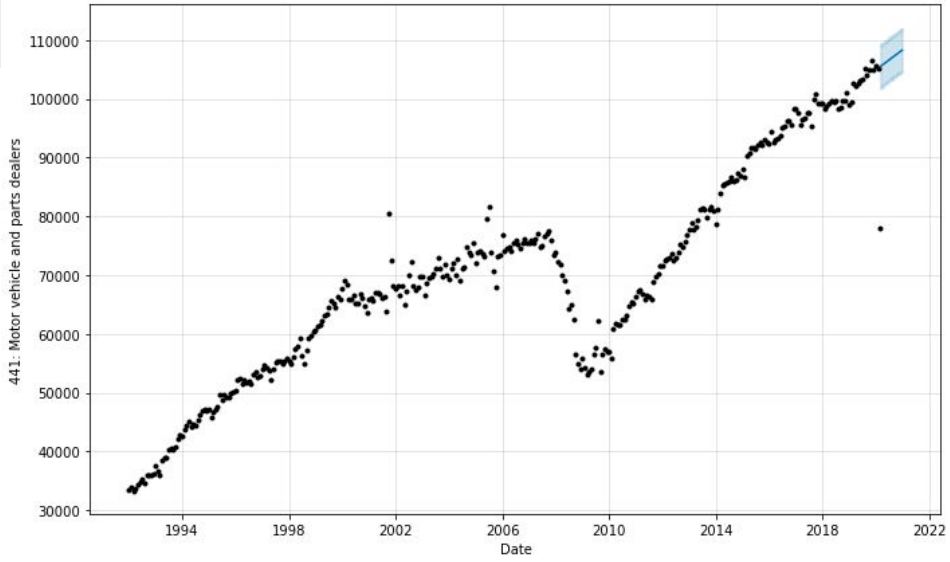
Conclusion

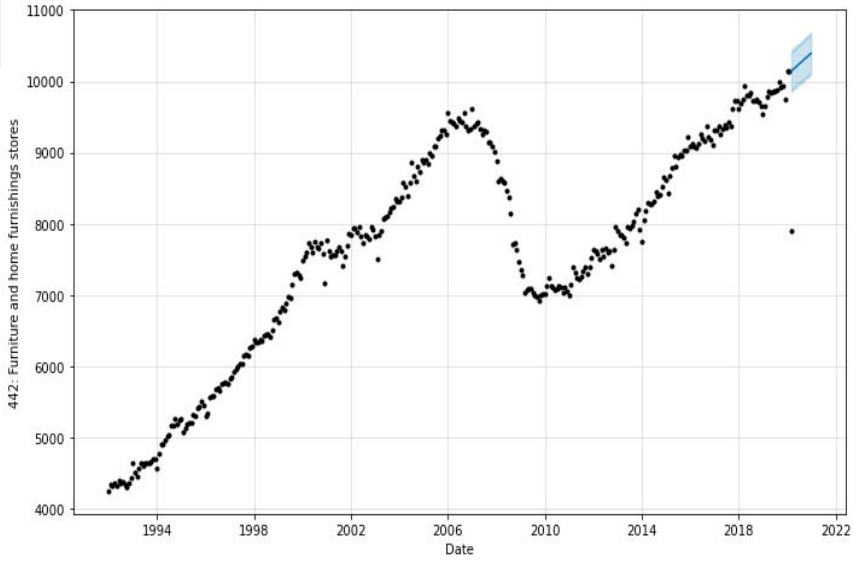
- comparing with the model fitted on non transformed data the standard error has significantly decreased
- The ACF doesn’t show significant autocorrelation among residuals.
- The red cure if from the model SARMA(p,q) ×(1,1)12 , which is very comparable to the original data which is shown by black line.
- These plot shows residuals points are now better at following normal distribution.
- The transformation does deal with the problem of the heteroscedasticity.

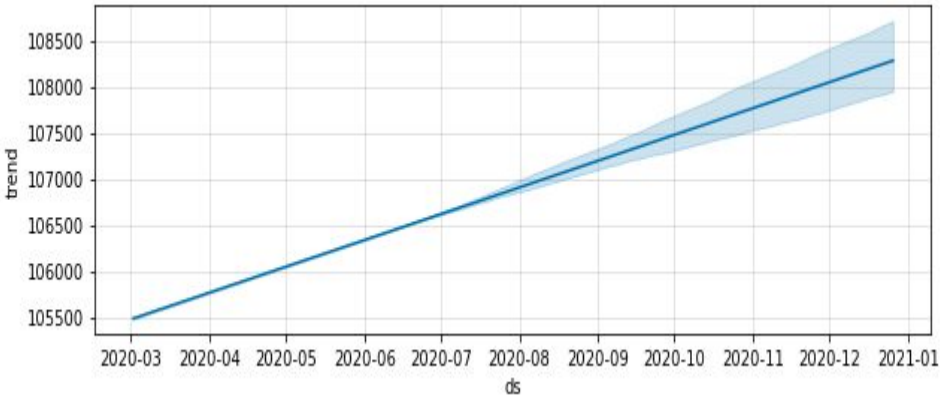
Facebook Prophet package

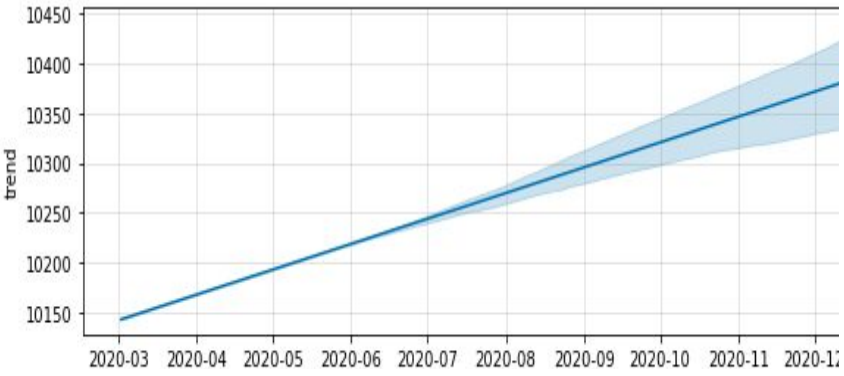
Facebook Prophet is a tool that allows folks to forecast using additive or component models relatively easily. It can also include things like:

- Day of week effects
- Day of year effects
- Holiday effects
- Trend trajectory
- Can do MCMC sampling


441: Motor vehicle and parts dealers
110000
100000
90000
80000
70000
60000
50000
40000
30000
1994
1998
2002
2006
2010
2014
2018
2022
Date


442: Furniture and home furnishings stores
11000
10000
9000
8000
7000
6000
5000
4000
1994
1998
2002
2006
2010
2014
2018
2022
Date


trend
108500
108000
107500
107000
106500
106000
105500
2020-03
2020-04
2020-05
2020-06
2020-07
2020-08
2020-09
2020-10
2020-11
2020-12
2021-01
ds


trend
10450
10400
10350
10300
10250
10200
10150
2020-03
2020-04
2020-05
2020-06
2020-07
2020-08
2020-09
2020-10
2020-11

Facebook Prophet package

- https://www.kaggle.com/landlord/us-retail-all-cat-covid-19-facebook-prophet
- Here is the complete forecasting of all catgiries using Facebook Prophet

Facebook Prophet package | Multiprocessing

- Adding multiprocessing to our code, Here we will launch a process for each time-serie forecast, so we can run our run_prophet function in parallel while we do the map of the list.
- https://www.kaggle.com/landlord/forecasting-multiple-time-series-using-prophet
- We could in the notebook see that using multiprocessing is a great way to forecasting multiple time-series faster, in many problems multiprocessing could help to reduce the execution time of our code.

LSTM for regression

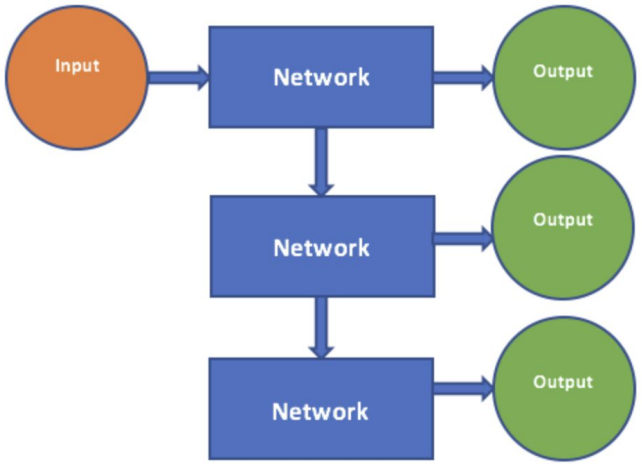
- Unlike regression predictive modeling, time series also adds the complexity of a sequence dependence among the input variables.
- A powerful type of neural network designed to handle sequence dependence is called recurrent neural networks. The Long Short-Term Memory network or LSTM network is a type of recurrent neural network used in deep learning because very large architectures can be successfully trained.

LSTM for regression

One to One. Classic Neural Network.

One to One

One to Many. Classic Neural Network.

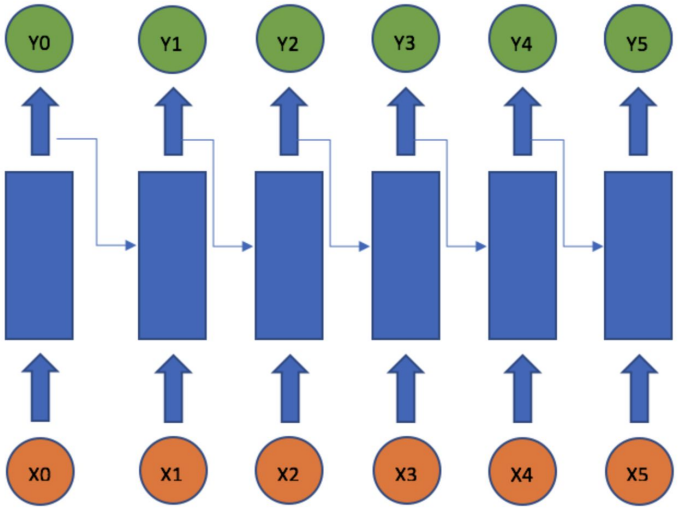


One to Many

Neelabh Pant. Last accessed October 15, 2018. <https://blog.statsbot.co/time-series-prediction-using-recurrent-neural-networks-lstms-807fa6ca7f>.

LSTM for regression

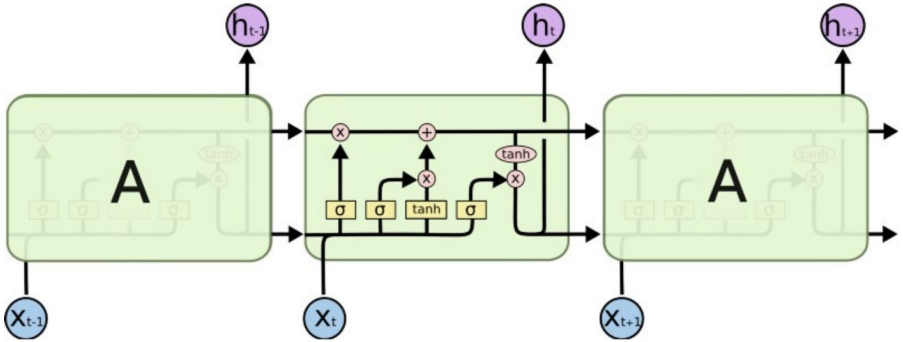
Recurrent Neural Network (RNN)



RNN Unrolled Time

$$Y_t = \tanh(wY_{t-1} + u\,x_t)$$

Long Short-Term Memory Network (LSTM)



LSTM Architecture

Able to capture longer-term dependencies in a sequence.

Neelabh Pant. Last accessed October 15, 2018. <https://blog.statsbot.co/time-series-prediction-using-recurrent-neural-networks-lstms-807fa6ca7f>.

LSTM for regression

Epoch 1/5 - 2s - loss: 0.0892

Epoch 2/5 - 1s - loss: 0.0268

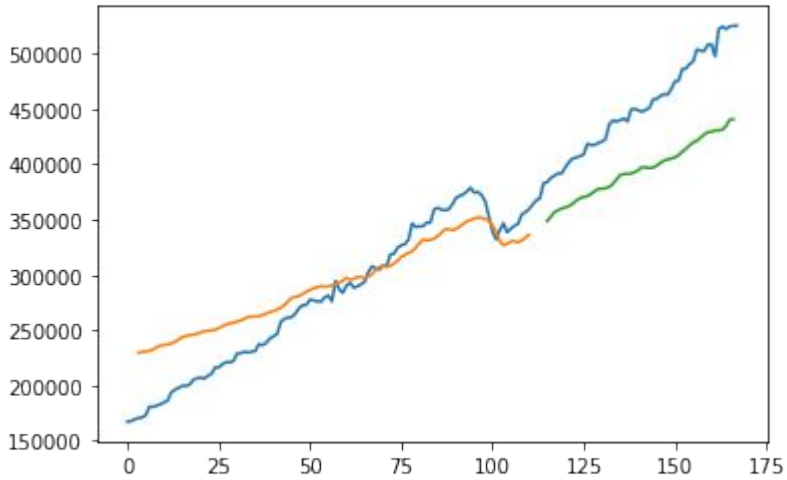
Epoch 3/5 - 1s - loss: 0.0169

Epoch 4/5 - 1s - loss: 0.0125

Epoch 5/5 - 1s - loss: 0.0080

Train Score: 27460.58 RMSE

Test Score: 60124.94 RMSE



You can see all working Ipython notebooks at:

Notebooks

You can see Dataset at:

Dataset

Thank you.
Vishwas Saini