

MACHINE LEARNING FORECAST MODEL IDENTIFYING LUXURY VEHICLE FLEET DEMAND



MARTELL TARDY, M.S.

DATA SCIENCE INTENSIVE CAPSTONE PROJECT

JUNE 30, 2021 COHORT

The Problem

Company

Lexus of Mishawaka is an authorized full-service Lexus dealership located in Mishawaka, Indiana offering all new and L/Certified Lexus vehicles and numerous luxury and mid-range vehicles from similar brands.

Context

In order to ensure Lexus of Mishawaka is accurately serving their market and the needs of their guests', the dealership has tasked their in-house Marketing and Information Specialist, Martell Tardy, with analyzing their historical sales data for insight.

Problem statement

How can the historical sales data from 2004 - 2017 be analyzed and used to deploy a machine learning model forecasting consumer demand in their market?

Challenges To Address

Challenge 1

Vehicle Sales History

The available dataset contains only sold vehicle information.

No vehicle information about the inventory sent to auction is available at this time.

Challenge 2

Record Keeping

The dataset contains inconsistencies in how sales information was recorded due to turnover and changes in technology overtime.

Challenge 3

Vehicle Identification

The Vehicle Identification Number (VIN) wasn't invented until 1954, therefore vehicles before this date sold in this dataset will not have information on vehicle specifications.

The Solution

Success for this project is the training and deployment of a machine learning model that can forecast which Lexus, Toyota, and non-Toyota models are necessary to have in the dealership inventory for the next 2 years starting April 2017.

This forecast will improve dealer order and inventory management, optimize plant production scheduling, and increase understanding of consumer demand in the market.

The Data

Data Information

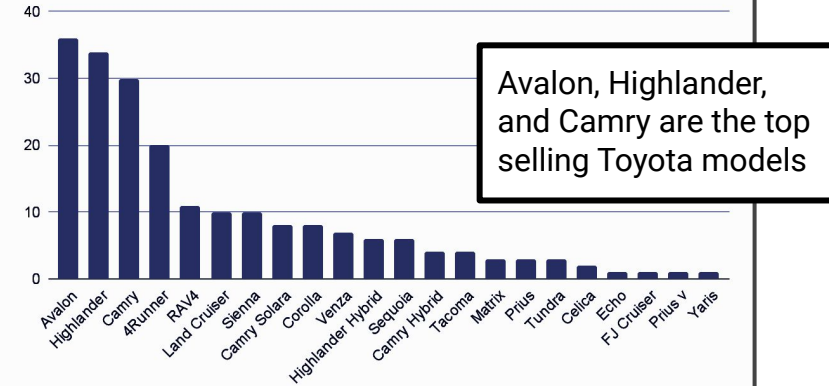
Data acquired: March 2017

For the period: June 2004 - March 2017

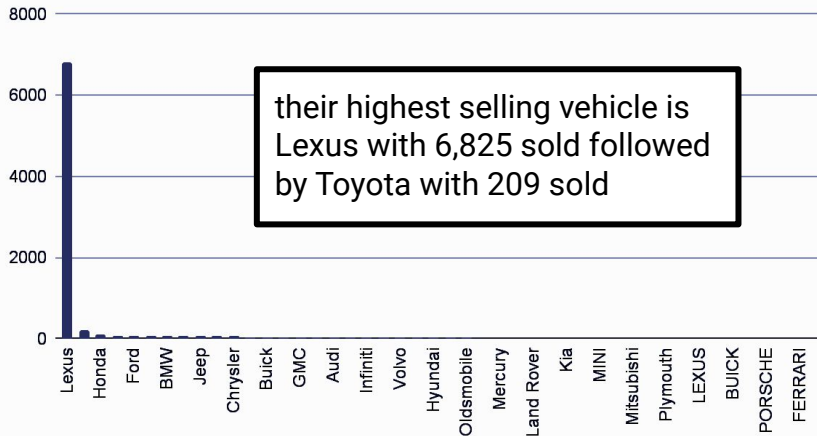
Number of records: 8,208

Number of fields: 34

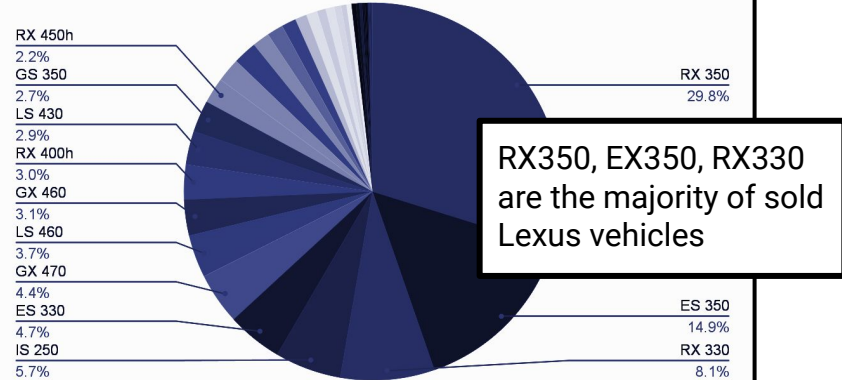
Toyota Sales By Model



Highest Vehicle Sales By Make



Highest Selling Lexus Models



Exploratory Data Analysis

Target Variables

	ContractYearMonth	TotalSales
0	2004-06	53
1	2004-07	53
2	2004-08	79
3	2004-09	64
4	2004-10	81
...
149	2016-11	37
150	2016-12	52
151	2017-01	36
152	2017-02	33
153	2017-03	45

154 rows x 2 columns

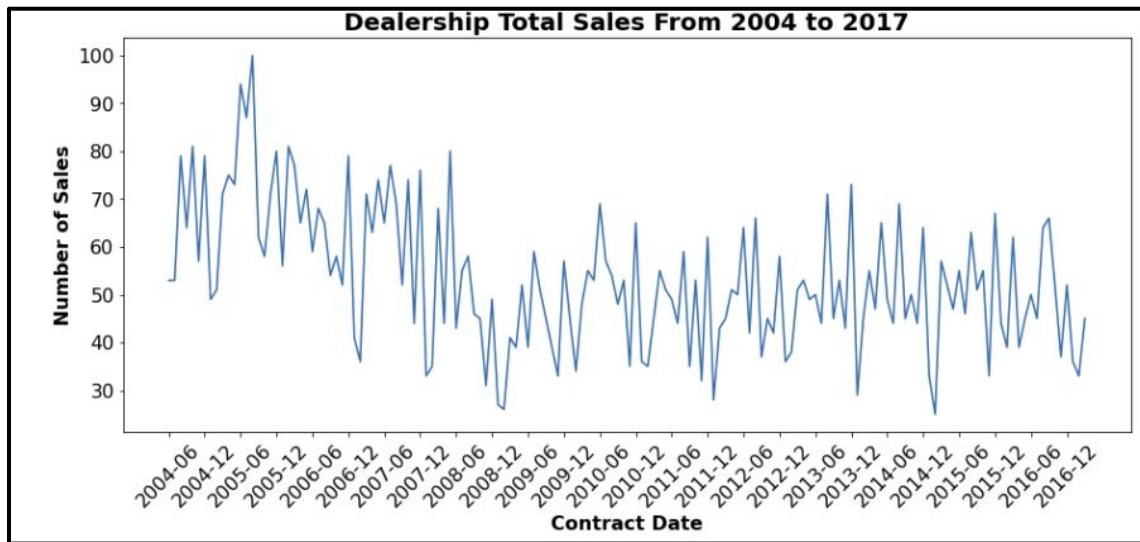
Two features were used to create the series time plot for this historical dataset.

The feature **ContractYearMonth** contains the timestamp of vehicle sales for a month in a specific year. The feature **TotalSales** contains the sum of sales for each of these respective timestamps.

This information was then filtered and explored as two separate datasets. The first dataset was named **df_total_sales** and contained sales history for all vehicle makes at the dealership. The second dataset was named **lex_sales** and contained sales for only Lexus vehicles sold by the dealership.

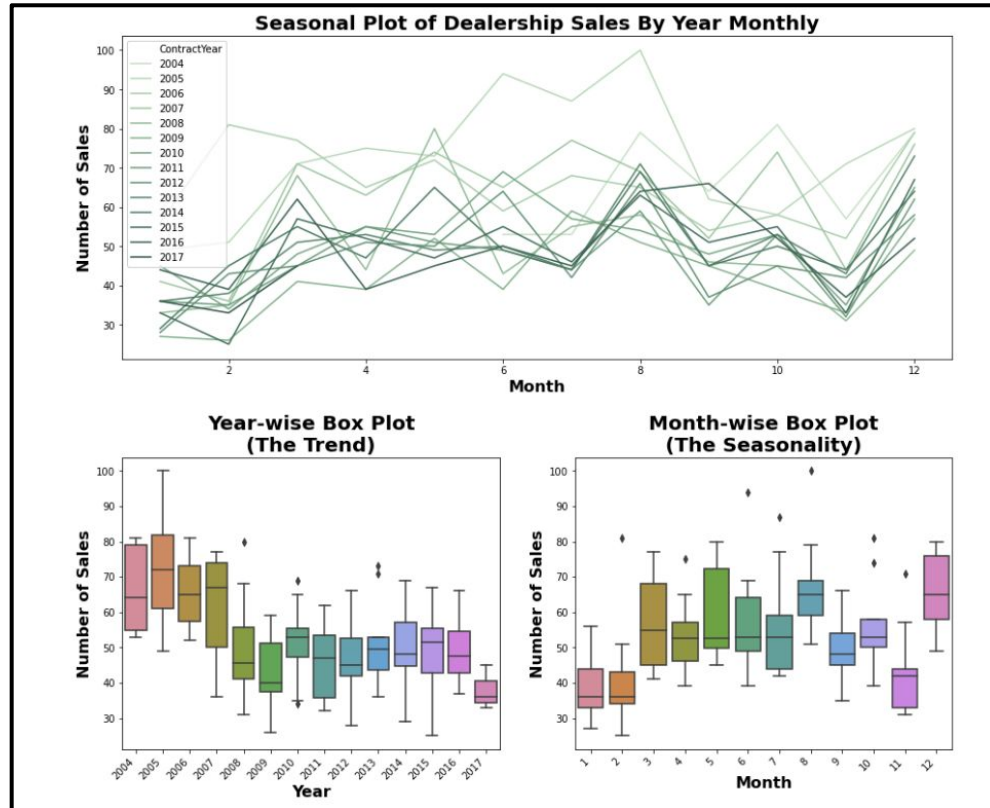
*no distinction between new or L/Cert classification was applied

Conversion & Time Plots



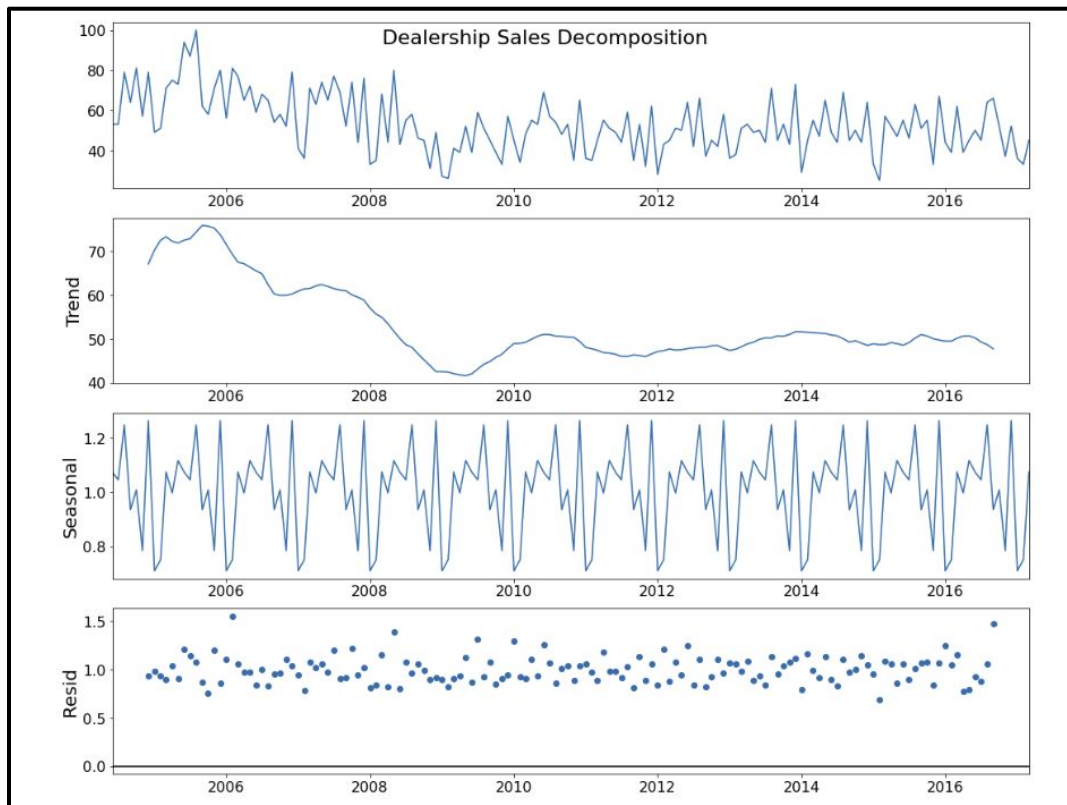
- Strong seasonality within each year
- Strong cyclic behavior over a period of 2–4 years
- Downward trend beginning mid to late 2007
- 2005 appears to have outliers and values which need to be explained
- Missing observations from January to May of 2004 and April of 2017 and onward
- Clear decreasing fluctuation in 2008, which is also during the last year of the financial crisis of 2007-2008

Seasonal Plots



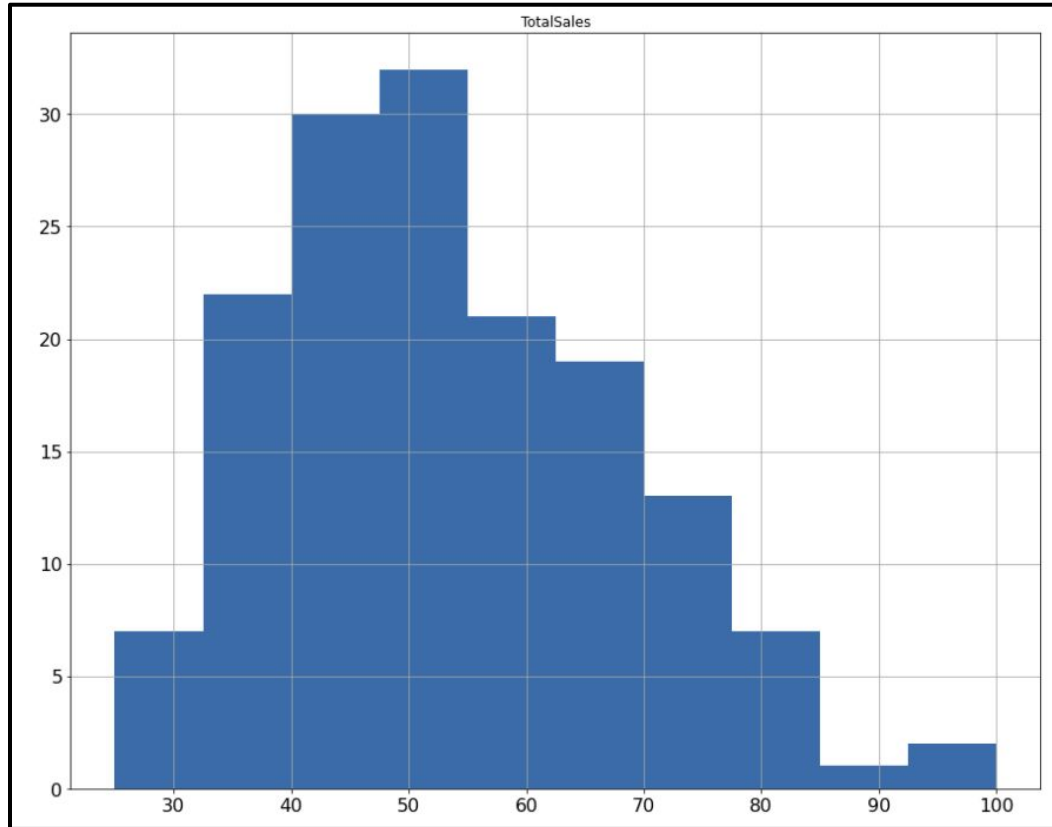
- Clear pattern of seasonal trend occurring every two months
- March, August, and December are typically in increasing fluctuation
- February, September, and November are always in decreasing fluctuation
- In the year-wise trend box plot clear outliers can be seen in 2008, 2010, and 2013
- In the month-wise seasonality box plot there are outliers in February(2), April(4), June(6), July(7), August(8), October(10), and November(11)
- clear downward trend yearly beginning in 2006 and a clear increasing fluctuation seasonally throughout the first two quarters of the year

Time Series Decomposition



- Downward trend beginning in late 2005 through 2009
- A climbing and dropping frequency that occurs every two years between 2010 through 2016, with 2010 having an increasing fluctuation
- Seasonal plot shows there is a pattern that occurs every two years
- Residual plot shows high variance in the early and late years of this time series.

Distribution of Observations



We can see from the histogram that the shape of the distribution of **TotalSales** observations yearly appears to be a right-skewed distribution. That is to say the mean is to the right of the median.

Therefore, we will need to transform the distribution of observations closer to a normal distribution.

To do so, we will use two methods and compare their results.

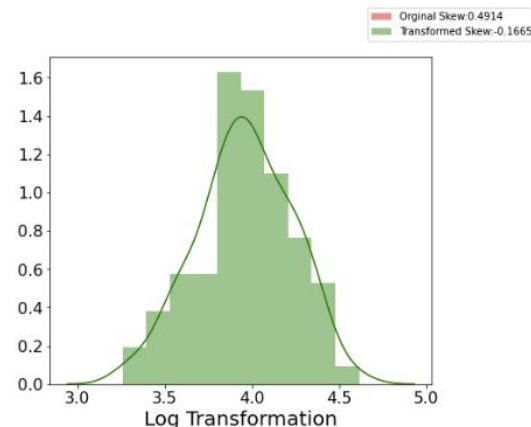
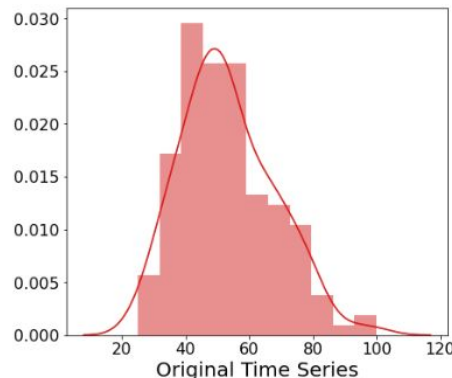
Transformations

Log Transformation

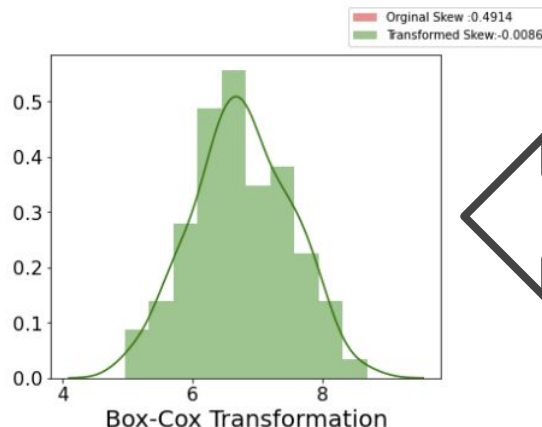
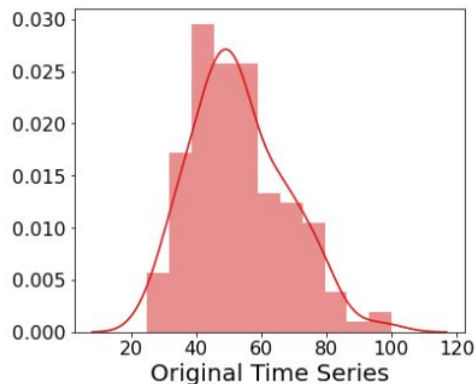
reduced skewness from

0.4914 to **-0.1665**

* Note, the best skew value should be nearly zero



The best lambda is 0.2519673728602007



Box-Cox Transformation reduced skewness from **0.4914** to **-0.0086**

The box-cox transformation has proven to have the best results and will be used for the post-hoc test

Year vs. Year

Tukey's Honestly Significant Difference Test

H₀ = no significant difference observed between the years

H_a = a significant difference has been observed between the years

Multiple Comparison of Means - Tukey HSD, FWER=0.05						
group1	group2	meandiff	p-adj	lower	upper	reject
2004	2005	0.2359	0.9	-0.783	1.2548	False
2004	2006	-0.0352	0.9	-1.0541	0.9837	False
2004	2007	-0.2389	0.9	-1.2578	0.78	False
2004	2008	-0.895	0.1514	-1.9139	0.1239	False
2004	2009	-1.2583	0.0034	-2.2771	-0.2394	True
2004	2010	-0.7303	0.4565	-1.7492	0.2886	False
2004	2011	-1.0076	0.0559	-2.0265	0.0113	False
2004	2012	-0.9426	0.1016	-1.9615	0.0763	False
2004	2013	-0.7783	0.3465	-1.7972	0.2406	False
2004	2014	-0.7816	0.3393	-1.8005	0.2373	False
2004	2015	-0.9043	0.1401	-1.9232	0.1146	False

As a result, we can now statistically confirm **2009** has the most statistically significant difference (negatively) in sales performance in comparison to **2004 through 2007**.

Also, **2017** has a notable statistically significant difference (negatively) in sales performance in comparison to **2004 through 2006** (whom have the highest sales performance, especially **2005**).

Lastly, **2013 and 2014** are the most statistically similar with -0.0033 mean difference. As well as, **2008 and 2015** with -0.0093 mean difference.

Handling Non-Stationarity

$p > 0.05$, therefore series is non-stationary

ADF Test

$p = 0.095768$

KPSS Test

$p = 0.010000$

$p < 0.05$, therefore again the series is confirmed as non-stationary

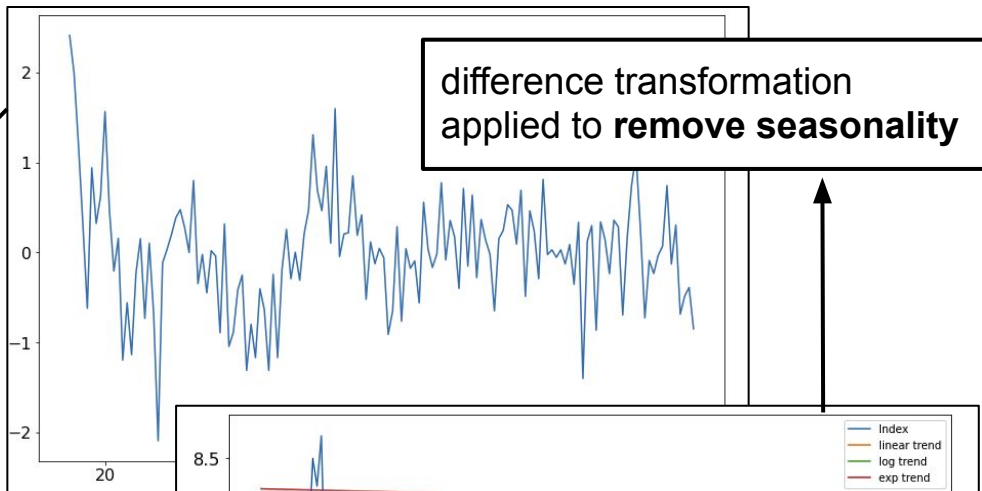
ADF Statistic: -4.258796
p-value: 0.000523
Critical Values:
1%: -3.483
5%: -2.884
10%: -2.579

final ADF Test to **confirm**
series is now stationary

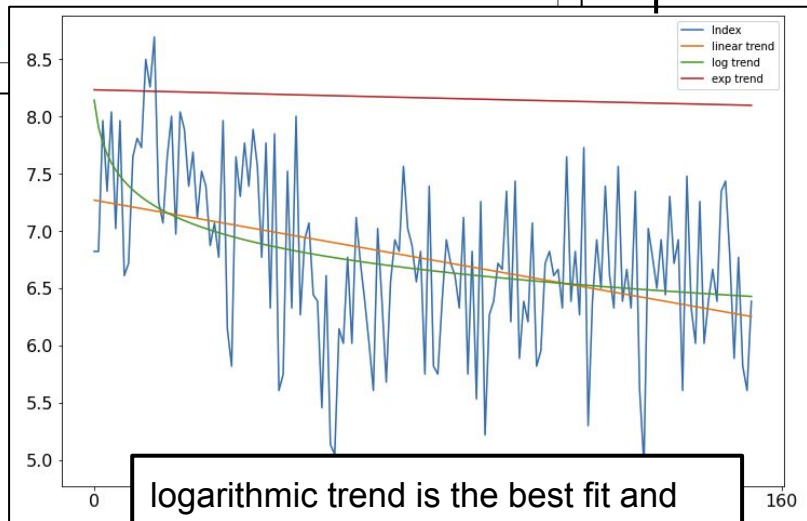
box-cox transformation
applied to the *TotalSales*
feature to **remove**
heteroskedasticity

Original Skew: 0.4914

Transformed Skew: -0.0086



difference transformation
applied to **remove seasonality**

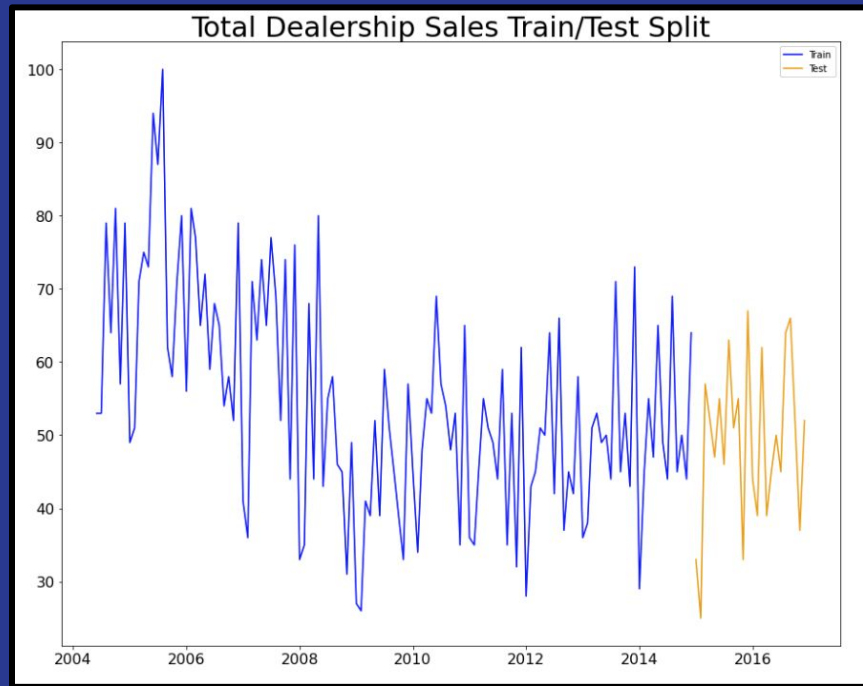


logarithmic trend is the best fit and
used to **remove trend** in the dataset

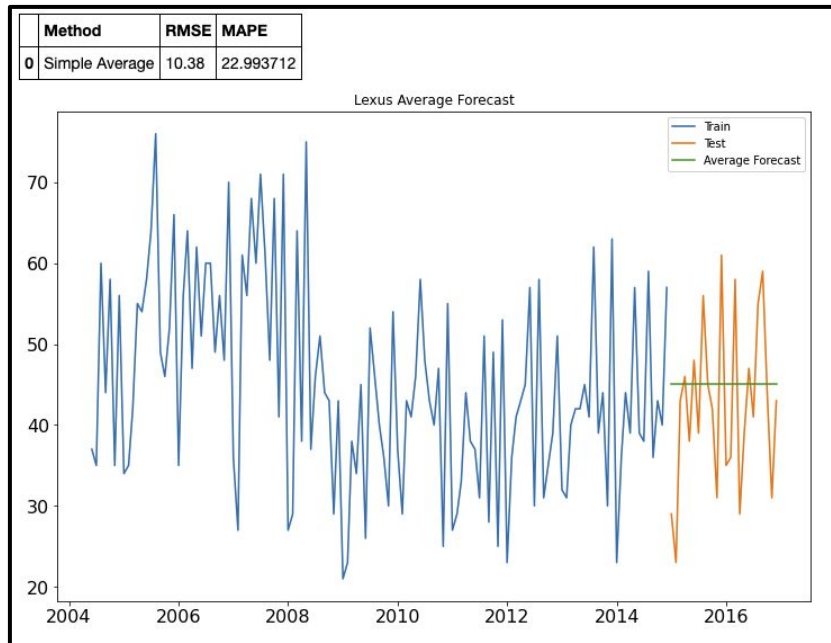
Preprocessing & Training

Train/Test Split

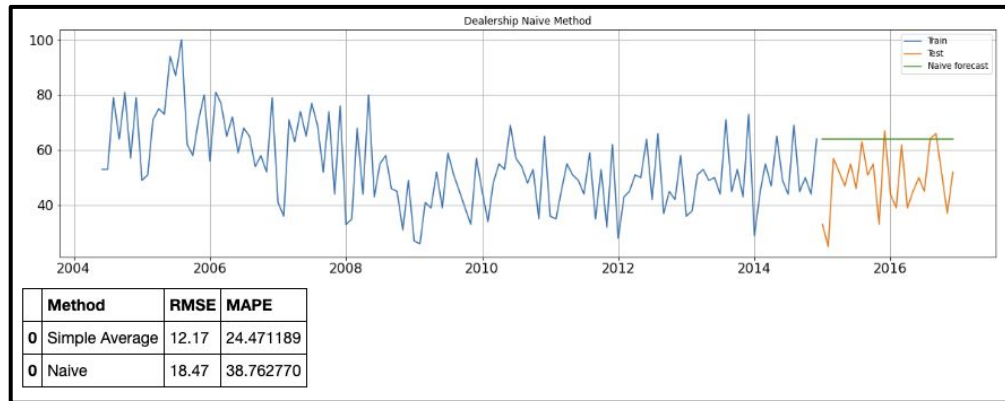
```
train = ts_dt['2004':'2014']  
test = ts_dt['2015':'2016']
```



Baseline Models: Simple Average & Naive Methods



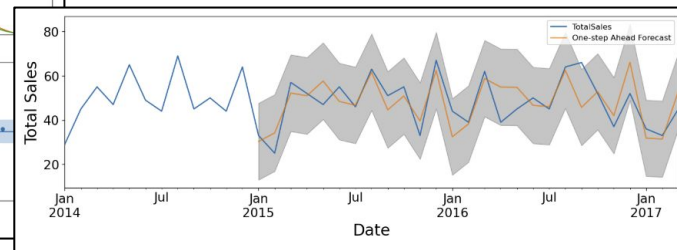
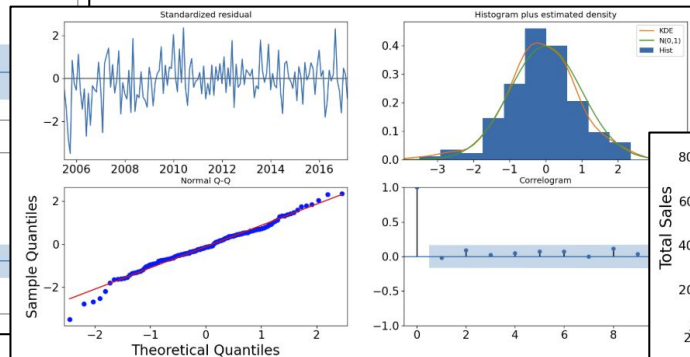
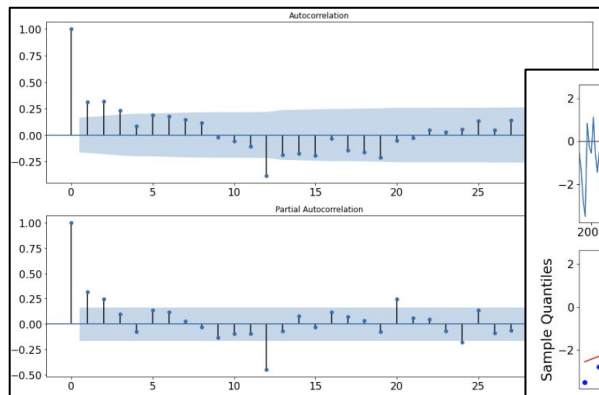
the **Simple Average** method forecasts of all future values are equal to the average (or “mean”) of the historical data



the **Naive** method is an estimating technique in which the last period's actuals are used as this period's forecast, without adjusting them or attempting to establish causal factors.

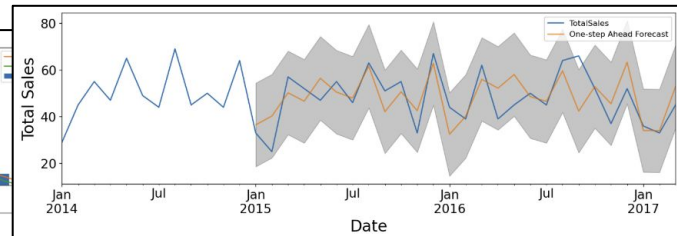
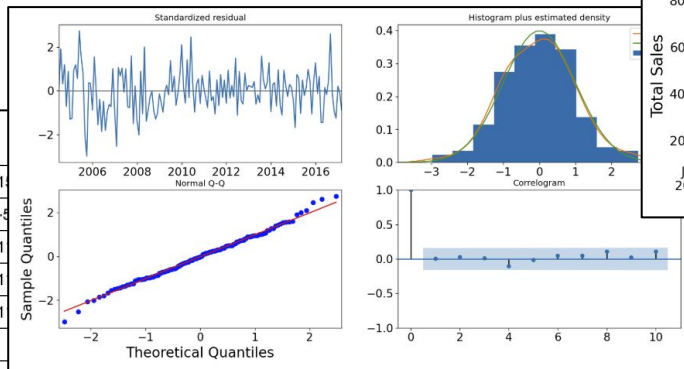
Baseline Models: SARIMAX Method

ACF & PACF
parameters
approach



Best model: ARIMA(2,1,1)(3,0,1)[12] intercept
Total fit time: 69.159 seconds

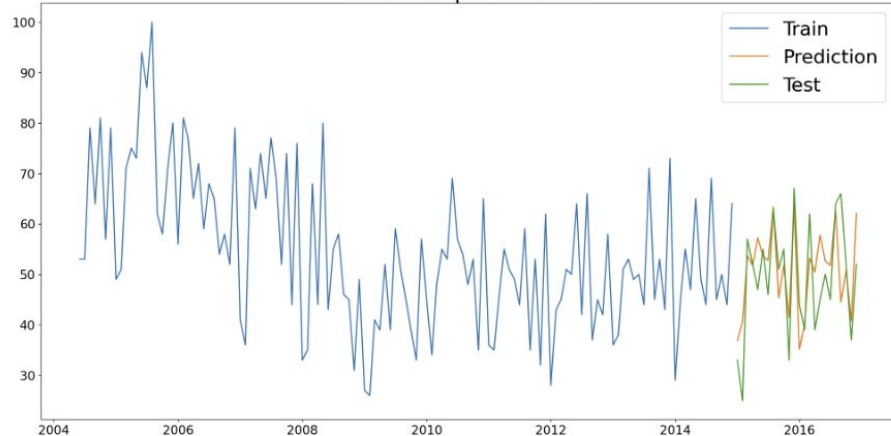
Dep. Variable:	y	No. Observations:	1
Model:	SARIMAX(2, 1, 1)x(3, 0, 1, 12)	Log Likelihood	-2
Date:	Sun, 27 Jun 2021	AIC	1
Time:	14:45:32	BIC	1
Sample:	0	HQIC	1
	- 154		
Covariance Type:	opg		



Grid Search
parameters approach

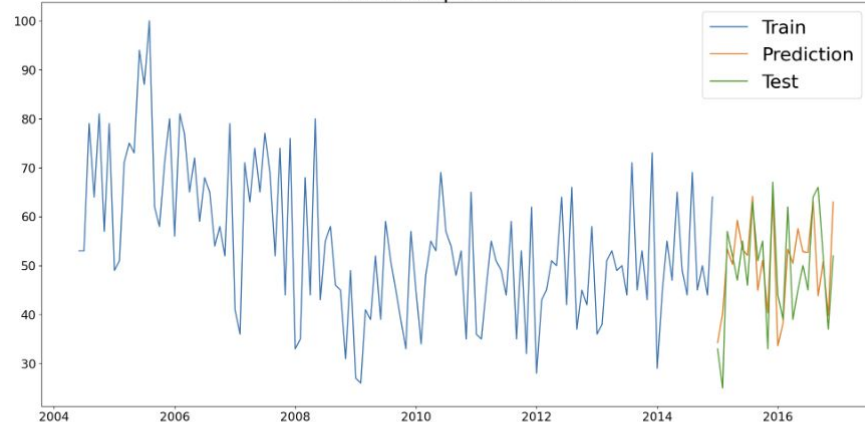
Baseline Models: SARIMAX Method (Continued)

Dealership Prediction



	Method	RMSE	MAPE
0	Simple Average	12.170000	24.471189
0	Naive	18.470000	38.762770
0	SARIMAX	8.241263	14.562545

Dealership Prediction



	Method	RMSE	MAPE
0	Simple Average	12.170000	24.471189
0	Naive	18.470000	38.762770
0	SARIMAX	8.241263	14.562545
0	SARIMAX_autoc	8.455708	14.561804

Grid Search parameter estimation approach suggested a SARIMAX model with the ordered parameters of $(2,1,1)(3,0,1)[12]$

ACF & PACF parameter estimation approach suggested a SARIMAX model with the ordered parameters of $(3,1,3),(1,1,1)[12]$

Baseline Models: Holt-Winter's Method

Grid Search Best Results

Trend: Multiplicative

Damped: False

Seasonal: Multiplicative

Seasonal Periods: 12

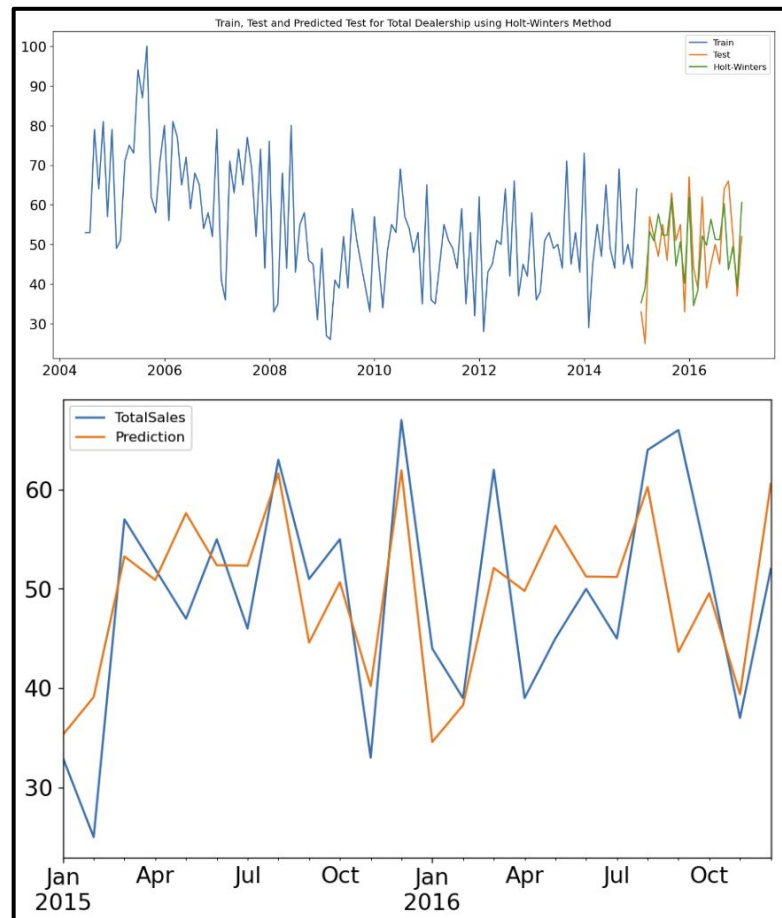
Box-Cox Transform: True

Remove Bias: False

*RMSE ~ 8.220 degrees

Validating Holt-Winter's
Predictions

Method	RMSE	MAPE
Holt-Winters	8.140466	14.206432



Model Evaluation

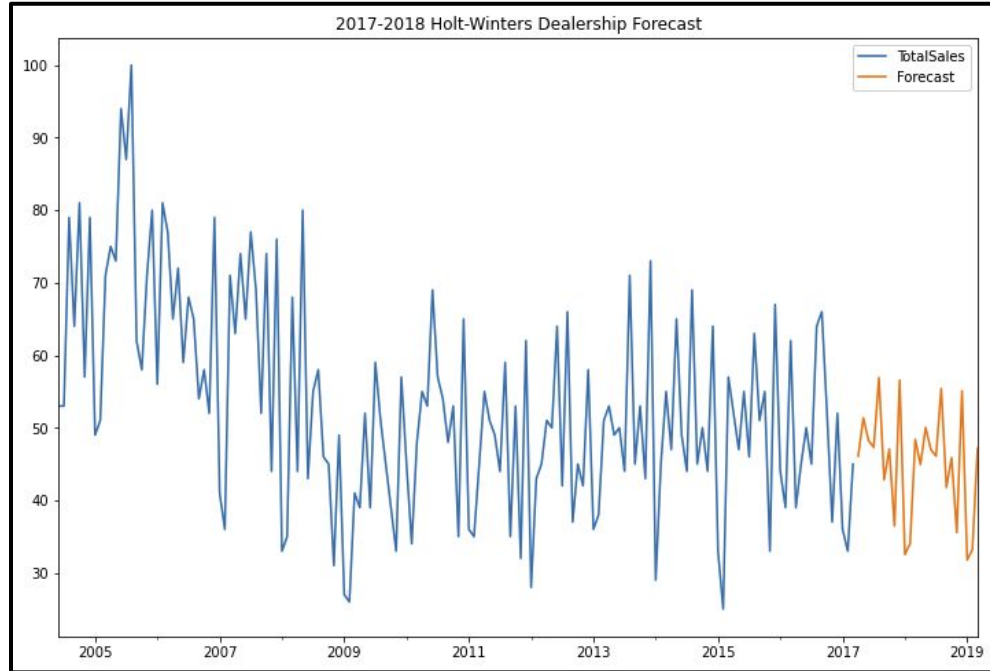
Out of the four baseline methods explored the **Holt-Winter's model performed the best with 85.8% accuracy** in predicting the test set observations

	Method	RMSE	MAPE
0	Simple Average	12.170000	24.471189
0	Naive	18.470000	38.762770
0	SARIMAX	8.241263	14.562545
0	SARIMAX_autoc	8.455708	14.561804
0	Holt-Winters	8.140466	14.206432



Machine Learning Model Forecast

Final Model Forecast: Holt-Winter's Method



24 months forecast of all vehicle sales for Lexus
of Mishawaka

2017-04-30	46.088141
2017-05-31	51.359114
2017-06-30	48.248633
2017-07-31	47.295832
2017-08-31	56.911114
2017-09-30	42.822225
2017-10-31	47.078531
2017-11-30	36.453334
2017-12-31	56.581796
2018-01-31	32.543081
2018-02-28	34.030010
2018-03-31	48.395210
2018-04-30	44.925273
2018-05-31	50.041999
2018-06-30	47.022739
2018-07-31	46.097768
2018-08-31	55.429885
2018-09-30	41.754070
2018-10-31	45.886807
2018-11-30	35.567699
2018-12-31	55.110345
2019-01-31	31.767856
2019-02-28	33.212963
2019-03-31	47.165030

A deeper look at
the forecasted
sales from **April**
2017 through
March 2019

Project Suggestions

It would be wise for the dealership to have a fleet from April 2017 through March 2019 that resembles a similar Lexus model selection to the units sold monthly historically. For example in April 2017, it would be strongly suggested that since the model forecast projected the total vehicle sales for the dealership to be around 46 units and about 38 of those to be Lexus vehicles, then as seen above, ~ 35% of those 38 Lexus vehicles should be the vehicle make RX350, ~ 20% ES350 and so on. This exact breakdown for April 2017 is recommended for April 2018 as well.
