Capstone Two

MACHINE LEARNING FORECAST MODEL
IDENTIFYING LUXURY VEHICLE FLEET DEMAND

MARTELL TARDY

06/29/2021

Report - IDENTIFYING LUXURY VEHICLE FLEET DEMAND

1. Introduction

1.1 Problem

Lexus of Mishawaka is an authorized Lexus dealership conveniently located on Grape Road in Mishawaka, Indiana. Lexus of Mishawaka is a full-service dealership offering their guests not only a full line-up of all new and L/Certified Lexus vehicles, numerous luxury and mid-range vehicles from similar brands, but also a friendly and reliable service and parts department. In order to ensure Lexus of Mishawaka is accurately serving their market and the needs of their guests', Lexus of Mishawaka has tasked their in-house Marketing and Information Specialist, Martell Tardy, with the task of analyzing their 2004 -2017 historical data for insight.

1.2 Criteria for Success

Success for this project would be the training and deployment of a machine learning model that will be able to forecast which Lexus, Toyota, and non-Toyota models are necessary to have in the dealership inventory 24 months 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.

1.3 Dataset

The dataset for this project was collected directly from the Lexus of Mishawaka Principle, Perry Watson III via their CRM system, VINSolutions in March 2017. The data consists of 8,208 entries of actual sales history which contains the following but not limited to information:

- Contract Date
- Vehicle Make
- Vehicle Model
- Contract Term
- Vehicle Sale Price
- Trade Vehicle Information

---

2. Data Wrangling

Prior to data wrangling, I removed all features pertaining to buyer contact information, such as buyer name and buyer phone number.

2.1 Historical Sales Data Dataset

The Historical Sales Data dataset contained 8208 entries of which there were 34 features 23 of which are categorical and 11 that are numerical, with 9 of those in float format.

Handling NaNs in Business Dataframe

Deleted Features (due to high null value count, or observations recorded show lack of relevance to scope of project, duplicate feature information, or impossible to replace null value accurately):

- Comments
- SalesPersonID
- SalesPersonID2
- BuyerBirthDate
- APR
- MonthlyPayment
- BuyerID
- StockNumber
- DeliveryDate
- DealStatus

Converted String (replaced with a string value "Not Applicable"):

- Trade1_StockNumber
- Trade1_VIN
- Trade1_Year
- Trade1_Make
- Trade1_Model
- Trade2_VIN
- Trade2_Year
- Trade2_Make
- Trade2_Model

Converted to a Value (converted after consulting owners of dataset, or compared to nearby relevant columns for context, or external web search for value):

- ContractTerm (1704 observations converted to '1.0' to represent cash buyers)
- FrontEndGrossProfit (700 observations converted to '0.00')
- BackEndGrossProfit
- BuyerHomeAddressPostalCode (12 observations converted to '11111' to represent true value unknown)
- BuyerHomeAddressState
- BuyerHomeAddressCity (13 observations converted to 'IN' for Indiana)
- VehicleSalePrice (1 observation converted to mean value of similar vehicles)
- InventoryType (9 observations converted to 'N' representing a new not used vehicle
- DealNumber (1 observation converted to generic number since a duplicate)

Summary of Business Dataset

1. There are 5979 unique VIN values observed, none of which are null values. 967 of them are repeated once, 97 twice, one three times, and one four times. It is suspected this is because these duplicated VINs belong to vehicles apart of lease programs.
2. The dealerships' top selling vehicle make is Lexus with Toyota at a far second followed by Honda, Mercedes-Benz, and Ford surprisingly competing for the consecutive places.
3. Lexus' top three selling vehicles are RX350, ES350 and RX330.
4. Toyota's top selling cars are the Avalon, Highlander, and Camry.
5. Removed 10 features reducing the dataset from 34 features to 23.
6. The transformed dataset is saved in the interim data folder of the project and now titled "Sales_Hist_Clean.csv".

---

3. Exploratory Data Analysis

3.1 Examining the target variable "ContractDate" and "TotalSales"

After completing the data wrangling of the Historical Sales Data dataset, it is now time to examine the target variable ContractDate and the total sales observed within this time period. Once a time plot is graphed it will be examined to see if there are consistent patterns? If there is a significant trend? If seasonality is important? If there is evidence of the presence of business cycles? If there are any outliers in the data that need to be explained by those with expert knowledge? How strong are the relationships among the variables available for analysis?

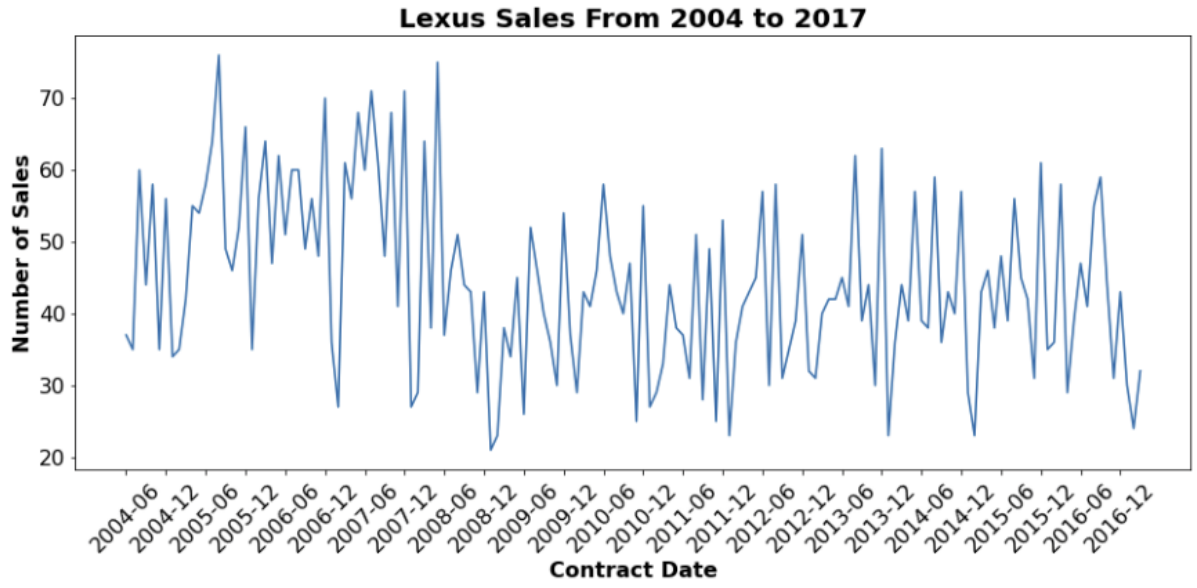
To properly explore the answer to these questions the Sales_Hist_Clean.csv dataset's ContractDate and total sales history will be examined using all vehicle sales and also using only the sales from Lexus vehicles.

3.2 Conversion & Time Plots

For time series data, the obvious graph to start with is a time plot. A time plot allows us to observe strange values like outliers, periods of missing observations, fluctuations within the data, and the potential trend, seasonality, and or cyclic behavior of the time series.

The time plot of the dealership's total sales shows strong seasonality within each year, as well as some strong cyclic behavior with a period of about 2–4 years. There is a downward trend in the data beginning mid to late 2007.

In addition, 2005 appears to have outliers and values which need to be explained because they differ from the seasonality or trend of any other year. There are missing observations from January to May of 2004 and April of 2017 and onward. Also there is a clear decreasing fluctuation in 2008, which is also during the last year of the financial crisis of 2007-2008 in the United States, and again in 2014.



From this time plot we can see the strong seasonality within each year, as seen in the dealership total sales time plot. The downward trend observed in the dealership total sales time plot is even more obvious when examining Lexus vehicle sales exclusively. The same

downward trend appears to begin in the first quarter of 2008 with it's lowest plunge at the start of 2009.

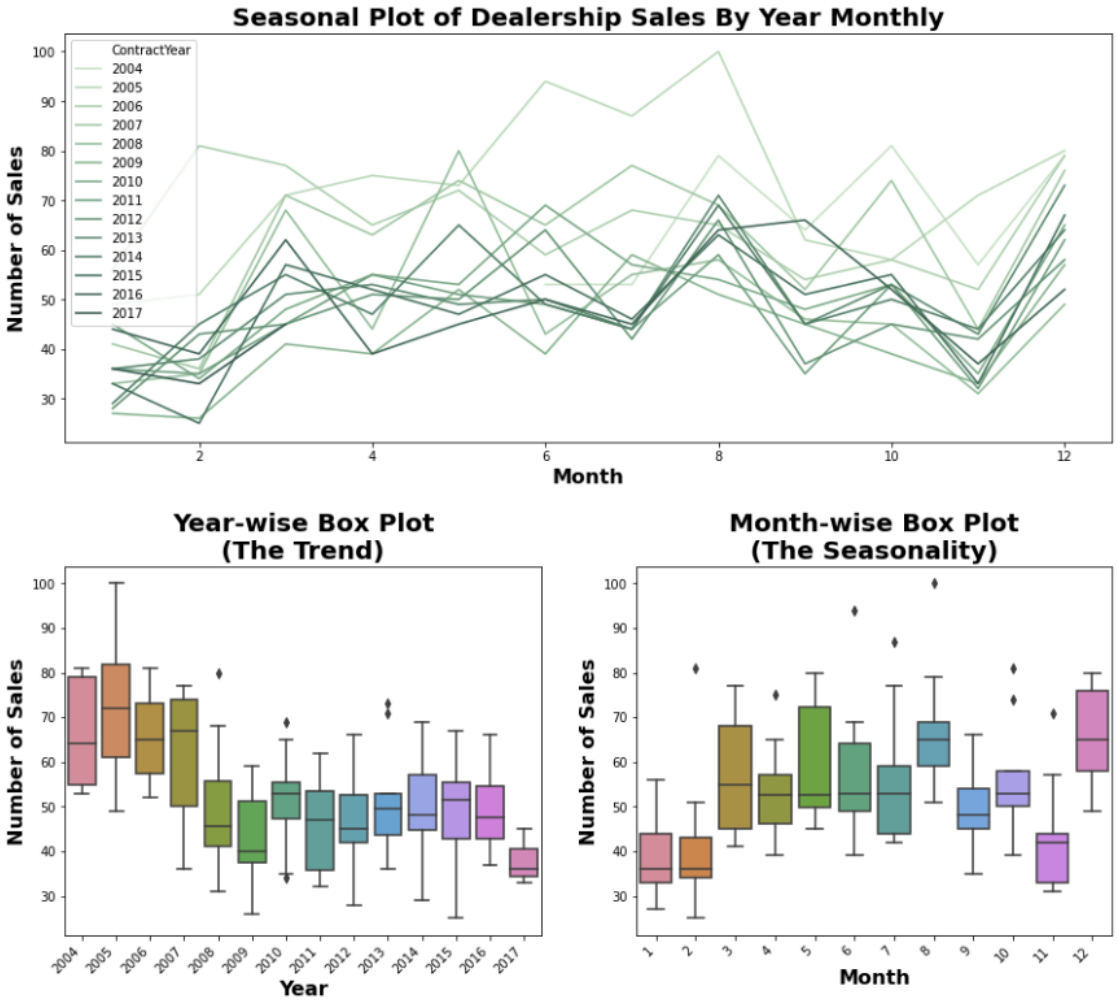
In addition, 2005 and 2006, as seen in the dealership total sales time plot, still appear to have outliers and values which need to be explained because they differ from the seasonality or trend of any other years. There are also missing observations from January to May of 2004 and the second quarter of 2017 and onward. Also there is a clear decreasing fluctuation in the third quarter of 2008, which coincides with the financial crisis of 2007-2008 in the United States.

3.3 Seasonal Plots

A seasonal plot is similar to a time plot except that the data are plotted against the individual “seasons” in which the data were observed. The data for each season are overlapped.

A seasonal plot consists of three parts: the seasonality plotted as a line graph, the trend behavior plotted yearly as a box plot, and the seasonality again plotted monthly but as a box plot.

Plotting the seasonality as a line graph allows us to observe more clearly the seasonal pattern if it exists monthly, the years in which the pattern changes, and any large jumps or drops in the time series. The trend and seasonality box plots allows the trend and the seasonality to be seen more clearly from a different perspective, observe if there are any years or months with outliers, and compare years or months easier if necessary.



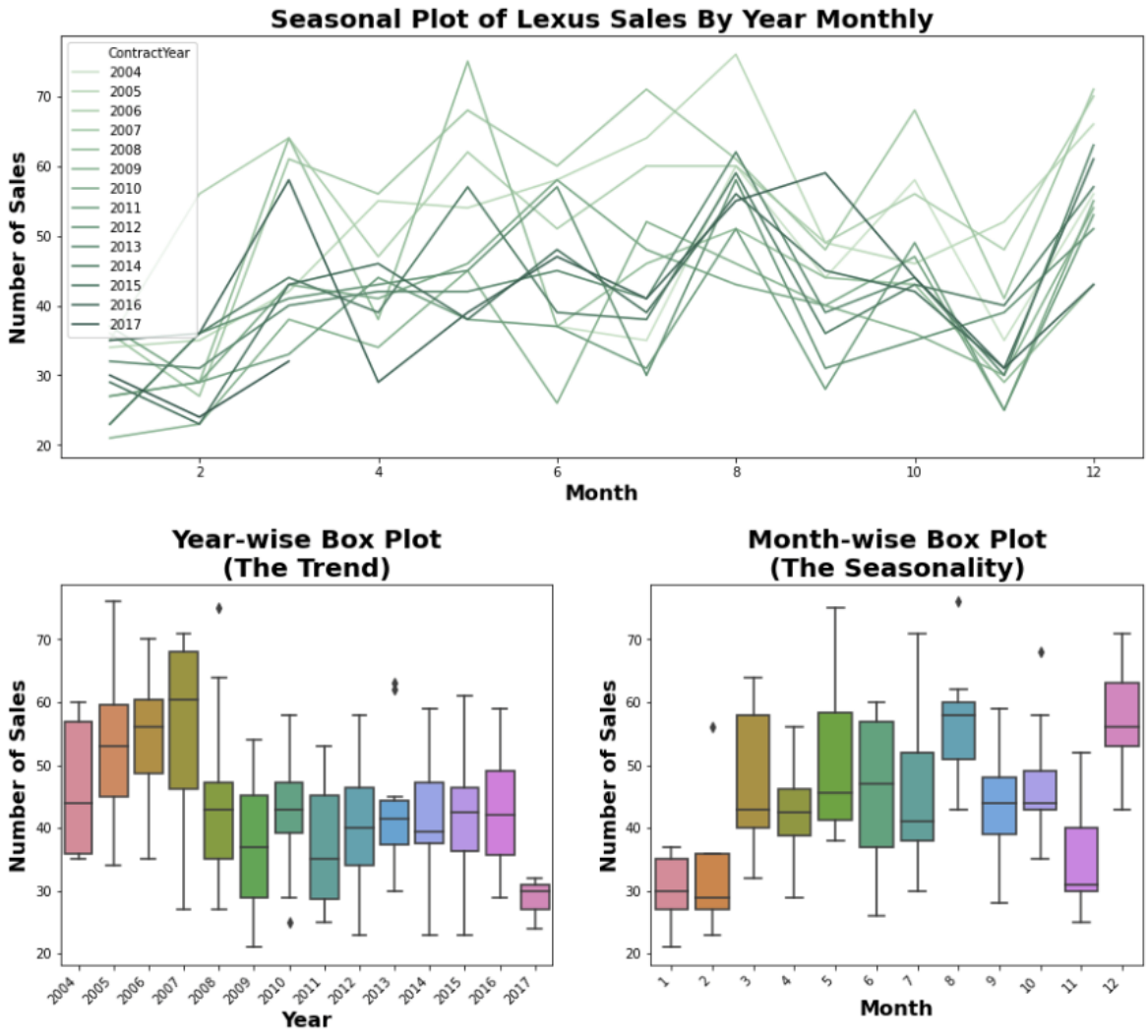
In the seasonal line plot there is a clear pattern of seasonal trend occurring every two months. March, August, and December are typically in increasing fluctuation, while February, September, November are always in decreasing fluctuation. As early as March Lexus dealerships begin to receive the new models for the year. August is the start of the new school year for many high schools and universities where a student vehicle is typical. Lastly, December is the month of Lexus' largest sales campaign, 'December to Remember'. These may be reasons why sales are typically in an increasing fluctuation during these months. February is a short month which could affect the number of selling days for the sales team and consumers are spending their money on Valentine's Day gifts. September is the first month after the summer months when people vacation and after the school year has begun. November is in the same month as one of the most expensive holidays, Thanksgiving. For these reasons this may be why a decreasing fluctuation is typical for these months.

In the year-wise trend box plot clear outliers can be seen in 2008, 2010, and 2013. Overall, it appears sales from 2004 to 2007 typically fall between 35 to 85 units sold, except for in 2005, where total sales exceed the typical range as observed in this cyclical pattern. 2008 is

the beginning of a new cyclical pattern that appears to last 2 years with a staggering decrease in total unit sells between 25 to the high 60s. Curious to know if the outlier in 2008 at 80 units was in the first quarter of the year. There is a increase in sales in 2010, but then a return to cyclical pattern started in 2008 with sales ranging between 25 to high 60s for the remaining 5 years of data.

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). January(1) and February(2) typically have the lowest total sales, with February as the lowest and December has the highest. From this plot, we now know the outlier we see in February occurred somewhere between 2004 to 2007, the outliers we see for June through August, belong to 2005, and the outliers for October and November occured between 2004 to 2007 or in 2013.

In conclusion, there is a clear downward trend yearly beginning in 2006 and a clear increasing fluctuation seasonally throughout the first two quarters of the year for the sale of all dealership vehicles.



In the seasonal line plot there is a clear pattern of seasonal trend occurring approximately every two months, except for June, which appeared to change from a decreasing fluctuation from 2004 onward until it becomes an increasing fluctuation from 2014 through 2016. The months of March, May, August, October, and December appear to have a consistent increasing fluctuation, while September and November appear to have a consistent decreasing fluctuation. As observed and stated for the dealership seasonal plot, as early as March Lexus dealerships begin to receive the new models for the year. August is the start of the new school year for many high schools and universities where a student vehicle is typical. Lastly, December is the month of Lexus' largest sales campaign, 'December to Remember'. These may be reasons why sales are typically in an increasing fluctuation during these months. February is a short month which could affect the number of selling days for the sales team and consumers are spending their money on Valentine's Day gifts. September is the first month after the summer months when people vacation and after the school year has begun. November is in the same month as one of the most expensive

holidays, Thanksgiving. For these reasons this may be why a decreasing fluctuation is typical for these months.

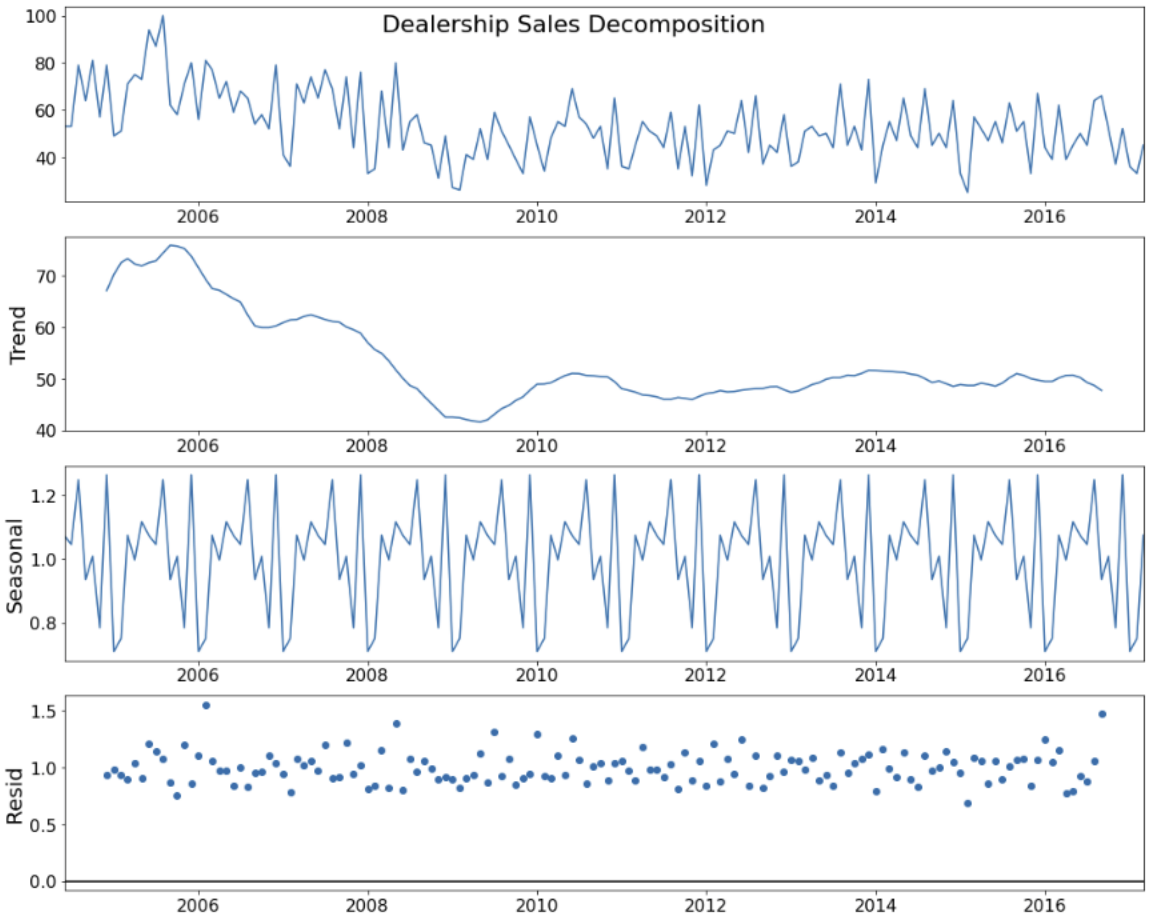
For the year-wise trend box plot clear outliers can be seen in 2008, 2010, and 2013. Overall, it appears sales from 2004 to 2007 typically fall between high 20s to low 70s units sold, except for in 2005, where total Lexus sales exceed the range, into high 70s, as observed in this cyclical pattern. 2008 is the beginning of a new cyclical pattern with a staggering decrease in total unit sales between low 20s to the low 60s. It can be observed as well that there is missing data for 2004 and 2017.

In the month-wise seasonality box plot there are outliers in February(2), August(8), and October(10), which is a significant decrease from the 7 identified for the dealership sales. January(1) and February(2) typically have the lowest sales, with May and December at typically the highest. From this plot, we now know the outlier we see in February could have occurred within any year except 2009, 2011, 2013, or 2017. The outlier we see in August, occured in 2005 or 2008, and lastly the outlier we see in October, could have occurred in 2005, 2006, or 2007.

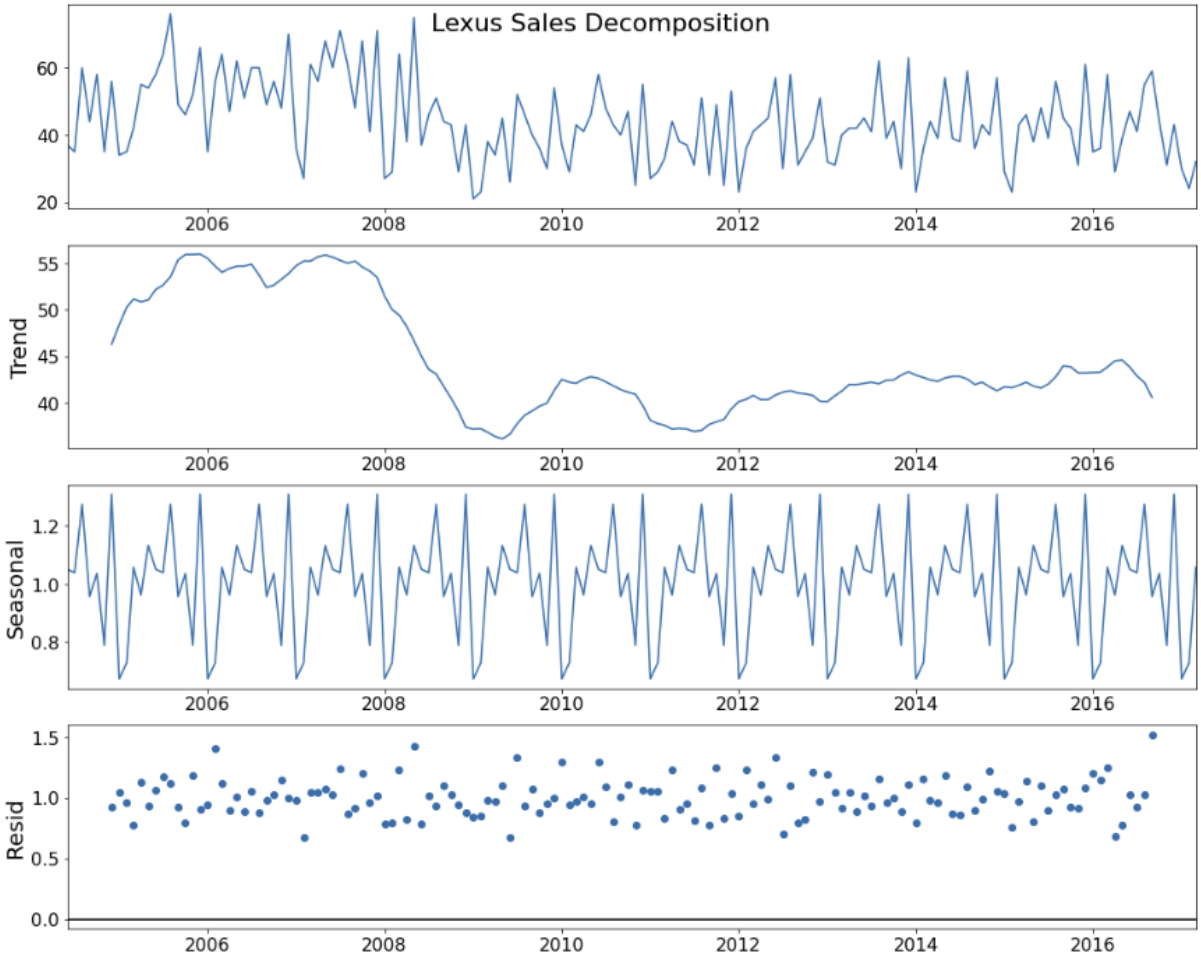
In conclusion, there is a clear downward trend yearly beginning in 2008 and a clear increasing fluctuation seasonally throughout the year for the sale of Lexus vehicles.

3.4 Time Series Decompositions

Since the historical sales data for the dealership and specifically their sale of Lexus units has a non-linear seasonality, meaning there were increasing and decreasing frequencies over time, a multiplicative model is appropriate, especially in the case of an economic time series. The seasonal_decompose() method from statsmodels library will be applied. This method decomposes a time series into trend, seasonality and noise.



In the first plot we see again the time plot created earlier in section 3.5.1.1 for the dealership's total sales. Next, is a plot of the trend. We can see a downward trend beginning in late 2005 through 2009. There is a climbing and dropping frequency that occurs every two years between 2010 through 2016, with 2010 having an increasing fluctuation. The seasonal plot shows there is a pattern that occurs every two years and the residual plot shows high variance in the early and late years of this time series.



In the first plot we see again the time plot created earlier in section 3.5.1.2 for Lexus' total sales displaying many highs and lows in the data. Next, is a plot of the trend, which unlike the dealership overall, Lexus appears to have an upward trend through 2005 and doesn't see a significant decline until the end of 2007. We can see this downward trend last through the first two quarters of 2009. Then a climb from the remaining half of 2009 and 2010, followed by a drop throughout 2011, then finally a very slow climb beginning in 2012 through the remainder of the data in 2016. The lack of data correlating to 2017, introduced a negative trend again, but this can't be confirmed. Notably, the trend never returns to the levels we see prior to 2007 which is similar to the dealership at large. The seasonal plot shows there is a pattern that occurs every year of significant increase followed immediately by significant decrease. Lastly, the residual plot shows high variance every two years, except in 2014 and very low variance in 2007, 2009, 2012, and 2016.

3.5 Examining Top Vehicle Makes and Models

Since we are analyzing the historical sales data from a Lexus dealership, exploring which vehicle brands and models make up the majority of sales is necessary in understanding the makeup of this dataset.

Lexus	6825
Toyota	209
Honda	96
Mercedes-Benz	87
Ford	84
Chevrolet	78
BMW	76
Cadillac	72
Jeep	71
Nissan	63

The top ten highest selling vehicles by make are listed above. Lexus makes up the majority of overall dealership sales. Number one is Lexus, which is the brand owner of the dealership where this data was collected from. In second place is Lexus' mothership company Toyota. From personal experience I know many Lexus owners began as Toyota owners and eventually became Lexus owners. The next three slots are close in sales, but Honda leads the pack. Understanding why Honda leads will not be explored in this notebook, but is interesting observation to explore further. Now, let's explore which vehicle models make up the majority of sales for this dealership.

RX 350	2033
ES 350	1016
RX 330	551
IS 250	388
ES 330	323
GX 470	302
LS 460	250
GX 460	209
RX 400h	204
LS 430	196

The top ten highest selling vehicle models for the dealership belong to the vehicle make Lexus. In first place is the RX350 which is the smallest SUV offered during this time period. Next, is the ES350 which is Lexus' mid-sized sedan. The next three positions belong to the RX SUV again and two sedans, IS250 and ES330, all of which are at the lowest trim level offered for that specific model.

3.6 Statistical Inference

The scientific question to be explored now is "Whether or not there's an association between the years provided in the historical sales data for the dealership and the number of units sold yearly?"

The hypothesis to test are: Ho = every year = same average number of sales, Ha = every year != same average number of sales

Since this dataset contains all recorded sales for this particular dealership during the given time period, we therefore have all recorded observations from the target population. As a result, there is no need for random sampling, which is used to make assumptions about the target population.

	TotalSales
ContractYear	
2004	66.571429
2005	72.583333
2006	65.500000
2007	61.833333
2008	48.916667
2009	42.333333
2010	51.333333
2011	46.333333
2012	47.583333
2013	50.500000
2014	50.500000
2015	48.666667
2016	49.583333
2017	38.000000

We can see 2005 was the strongest year for sales with an average of 72.5 sales and 2017 was the weakest year for sales with an average of 38 sales. The question still remains, are the differences among the years due to statistically significant differences among the population mean or merely due to insignificant variability? One-Way ANOVA F-test will assist in answering this question.

3.7 One-Way ANOVA F-test

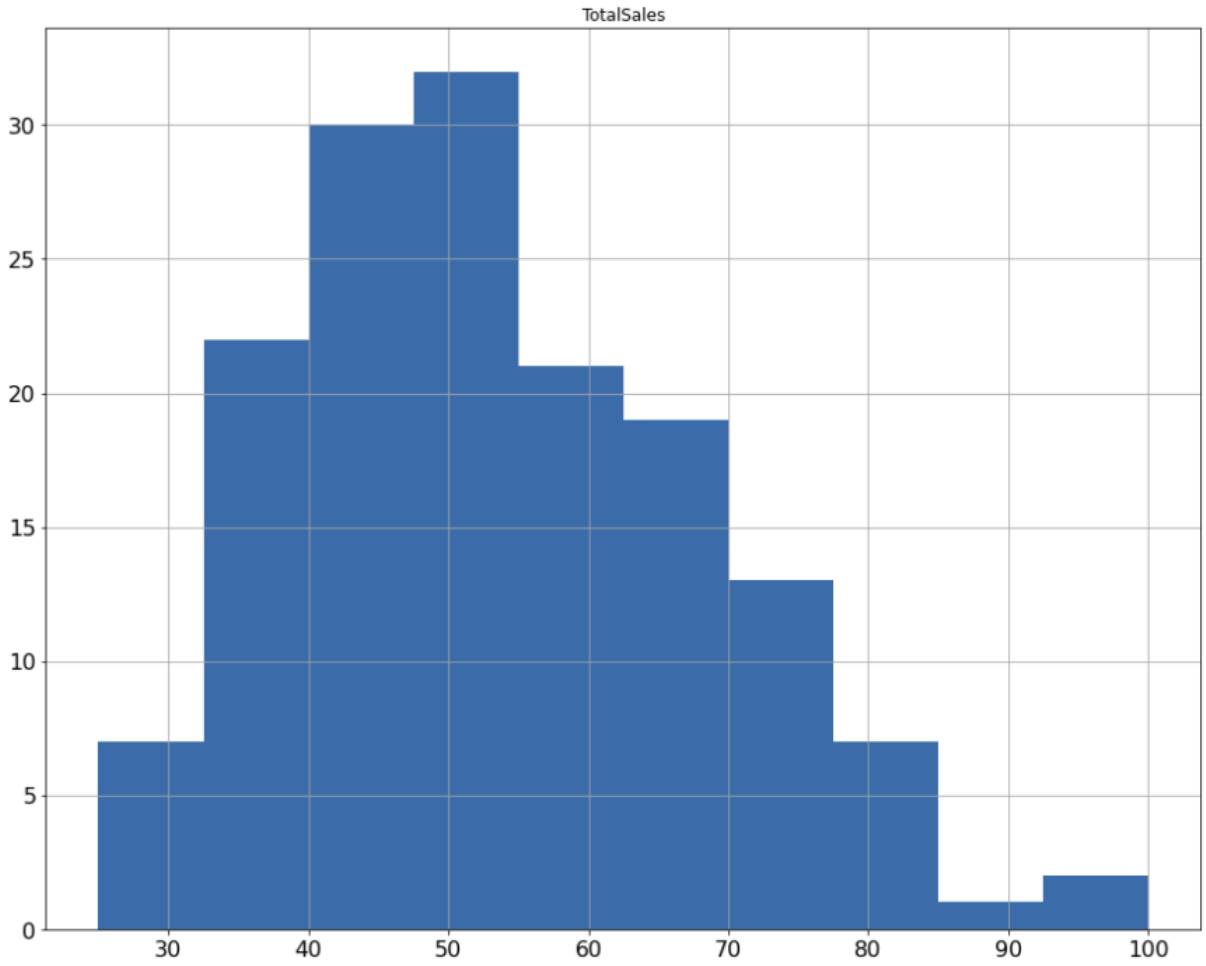
	df	sum_sq	mean_sq	F	PR(>F)
ContractYear	13.0	12363.545455	951.041958	6.556096	1.197142e-09
Residual	140.0	20308.714286	145.062245	NaN	NaN

The null hypothesis states that the mean of all groups is equal, which implies that our model has no explanatory value and that there is no proof for choosing one year over another. The alternative hypothesis states that at least one of the means is different, which would be a reason to go more in-depth and find out which year is weakest or strongest and why.

Our p-value of 1.197142e-09 as a real number is 0.000000001197142. This is significantly lower than 0.05, so we can reject our null hypothesis and accept our alternative hypothesis: stating that at least one of the means is different and that there is an association between the year of sales and the number units sold yearly. To answer statistically which of the years differ the most in sales performance and which are similar, a post-hoc comparison test will be performed.

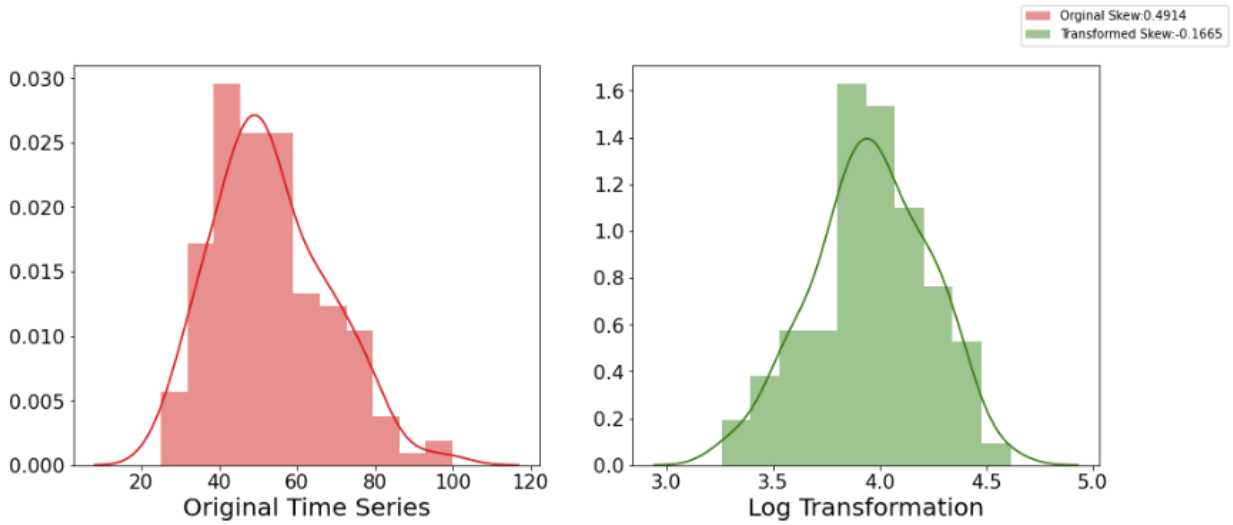
Post-hoc tests assume the distribution of observations is a fairly normal distribution. To confirm if this is true for the dataset, we will use a histogram to visualize the distribution and then perform any necessary transformations.

Exploring Distribution of Observations



We can see from the histogram that the shape of the distribution of observations 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.

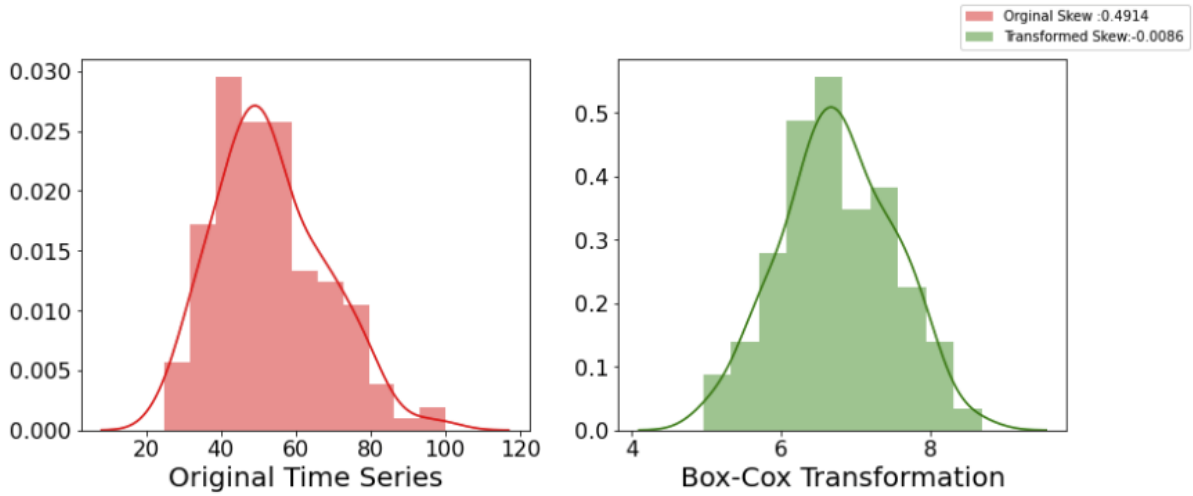
Log Transformation



The above plots are the comparison of the original distribution and the log transformation of the distribution. Log transformations are effective at removing exponential variance from time series data and we can see this through the reduction in skewness from 0.4914 (red) to -0.1665 (green). Note, the best skew value should be nearly zero. Now, let's compare the results of a box-cox transformation.

Box-Cox Transformation

```
The best lambda is 0.2519673728602007
```



The above plots are the comparison of the original distribution and the box-cox transformation of the distribution. Box-Cox transformation considers all the values of lambda (from -5 to 5) and the best value for the data is selected. The “Best” value is one that results in the best skewness of the distribution. From the print output we can see the best lambda is 0.2519673728602007, which is between a log transform lamda value of 0.0 and a square root transform lambda value of 0.5. In conclusion, we can see from the legend that the original skew has been reduced from 0.4914 (red) to -0.0086 (green) with the box-cox

transformation. As a result, the box-cox transformation of the dataset has proven to have the best results and therefore will be used as the transformed dataset for our post-hoc test.

3.8 Post-Hoc Test: Tukey's Honestly Significant Difference Test (Tukey HSD)

When the explanatory variable is more than two groups, a significant ANOVA does not tell us which groups are different from the others. This is the case for this time series.
The Tukey's Test is one of the most commonly used post-hoc tests, which allows pairwise comparisons between the means of each group while controlling for the family-wise error rate. This method tests at $P<0.05$ (correcting for the fact that multiple comparisons are being made which would normally increase the probability of a significant difference being identified).

For this project, each year within the dataset will be considered a separate group (explanatory variables) and the total number of vehicles sold for each corresponding year will be the response variable. The difference observed between any given two years will be stated as "False", meaning the acceptance of the null hypothesis or "True" meaning the rejection of the null hypothesis.

The hypotheses for this test are the following: Ho = there is no significant difference observed; Ha = a significant difference has been observed.

(visual of output too large to display)

As a result of the Tukey HSD Test we can see the differences and similarities between all the years within this time series. Notably are the "True" outputs observed within the "reject" column, which state there is a significant difference observed between the correlating group1 and group2. In addition, the column "meandiff" states the mean difference between the two groups.

In this regard, when 2004 is put in contrast to 2009 and 2017 these two years are "True" for the rejection of the null hypothesis and the significant difference observed is a negative mean difference. Therefore, 2009 and 2017 are years that have a significant decline in sales when compared to sales observed in 2004.

Then when 2005 was examined in comparison to 2008 through 2017, they also reject the null hypothesis, with an output of "True", and the significant difference observed is also a negative mean difference. It's to be noted that 2005 when compared with 2006 and 2007 also show a negative mean difference, but with a p-value adjusted ("p-adj") greater than 0.05. Therefore, the null hypothesis cannot be rejected in these cases. The only positive mean difference observed between 2005 and another year is 2004. This observations make sense, since 2005 has the highest number of total sales within this time series.

Next, 2006 in comparison to 2009, 2011, 2012 and 2017, all reject the null hypothesis as well for the same reasons stated. Finally, 2007 in comparison to 2009 rejects the null hypothesis as well.

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

This is not to say 2009 has the lowest number of sales in this time series, that is 2017, but rather it is the most frequent year to have a significant difference (negatively) when compared to 2004 through 2007.

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.

4. Augmented Dickey Fuller (ADF) Test

Dealership ADF Test:

```
ADF Statistic: -2.586742
p-value: 0.095768
Critical Values:
        1%: -3.478
        5%: -2.882
        10%: -2.578
```

Since the p-value = 0.095768 is greater than the significance level [0.05], therefore the null hypothesis cannot be rejected and the time series appears non-stationary. Let's run a second test to confirm that the time series is non-stationarity.

Lexus ADF Test:

```
ADF Statistic: -1.786583
p-value: 0.387143
Critical Values:
        1%: -3.478
        5%: -2.883
        10%: -2.578
```

Since the p-value = 0.387143 is greater than the significance level [0.05], therefore the null hypothesis cannot be rejected and the time series appears also non-stationary.

Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test

KPSS test is another test for checking the stationarity of a time series. The null and alternative hypothesis for the KPSS test are opposite that of the ADF test. **Ho: The process is trend stationary. Ha: The series has a unit root (series is not stationary).

The cleaned dataset is used for statistical analysis and machine learning. It contains 39,824 rows with 158 columns. All columns have float64 data-type ranging from 0 to 1.

Dealership KPSS Test:

```
Results of KPSS Test:
Test Statistic           1.007066
p-value                  0.010000
Lags Used                7.000000
Critical Value (10%)     0.347000
Critical Value (5%)      0.463000
Critical Value (2.5%)    0.574000
Critical Value (1%)      0.739000
```

Based upon the significance level of 0.05 and the p-value of KPSS test (0.010000), there is evidence for rejecting the null hypothesis in favor of the alternative. Hence, the series is non-stationary as per the KPSS test.

Lexus KPSS Test:

Based upon the significance level of 0.05 and the p-value of KPSS test (0.010000), there is evidence for rejecting the null hypothesis in favor of the alternative. Hence, the series is non-stationary as per the KPSS test.

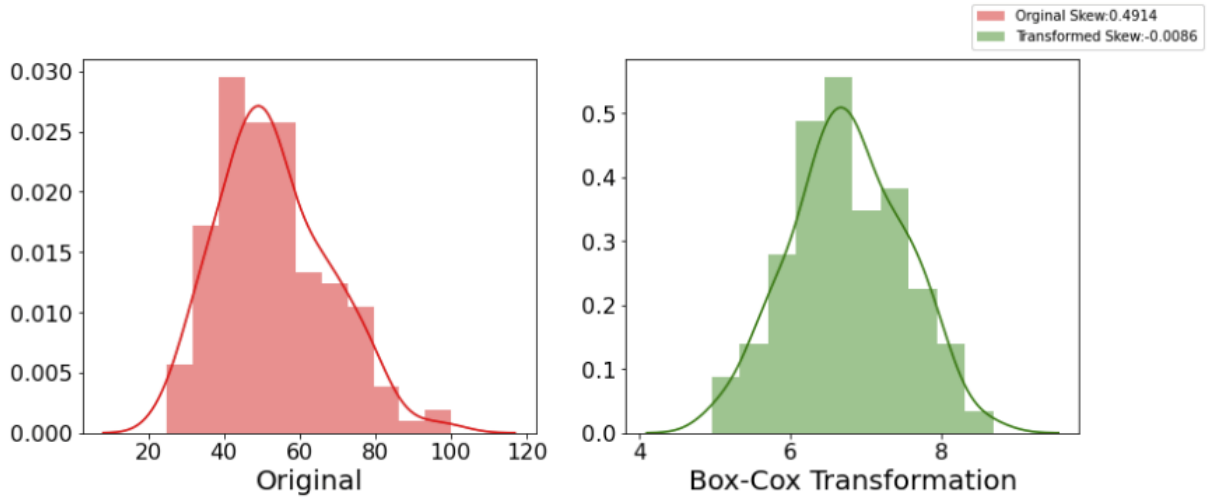
It is always better to apply both the tests, so that it can be ensured that the series is truly stationary. Now, we will move forward with finding the order of differencing.

Handling Heteroskedasticity

To make the time series stationary, we need to apply transformations to it. Since box-cox transformation was applied on the yearly distribution and significantly reduced skewness, the same transformation will be applied to the dealership (df = df_total_sales['TotalSales']) and the Lexus (df = df_total_sales['TotalSales']) datasets to remove heteroskedasticity.

Dealership Box-Cox Transformation

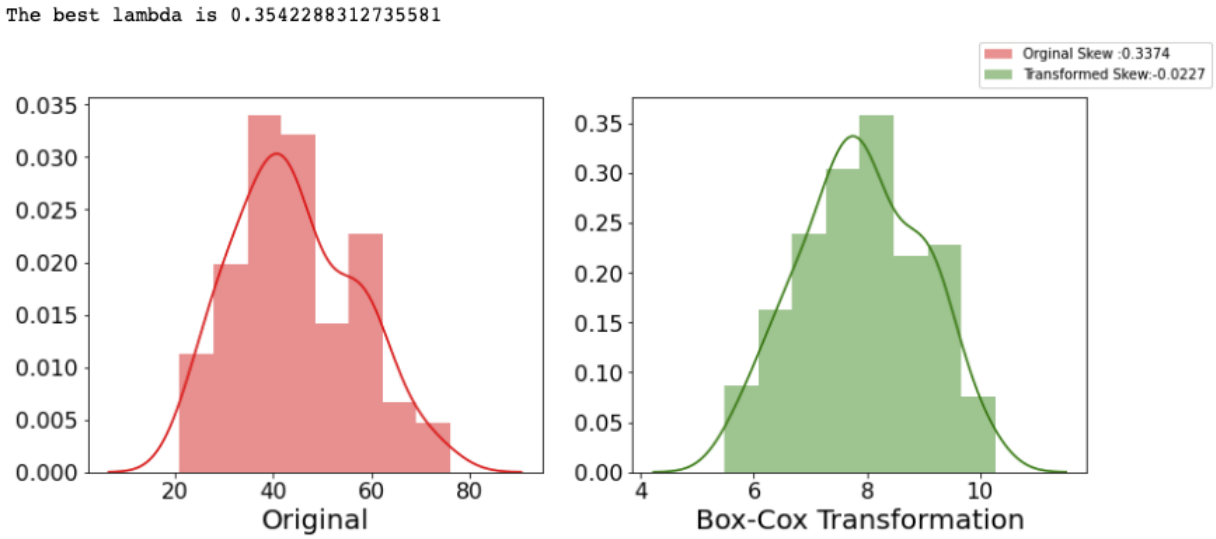
```
The best lambda is 0.2519673728602007
```



From the print output we can see the best lambda is 0.2519673728602007 again, which is between a log transform lamda value of 0.0 and a square root transform lambda value of 0.5.

In conclusion, we can see from the legend that the original skew has been reduced from 0.4914 (red) to -0.0086 (green) with the box-cox transformation and therefore handling heteroskedasticity.

Lexus Box-Cox Transformation



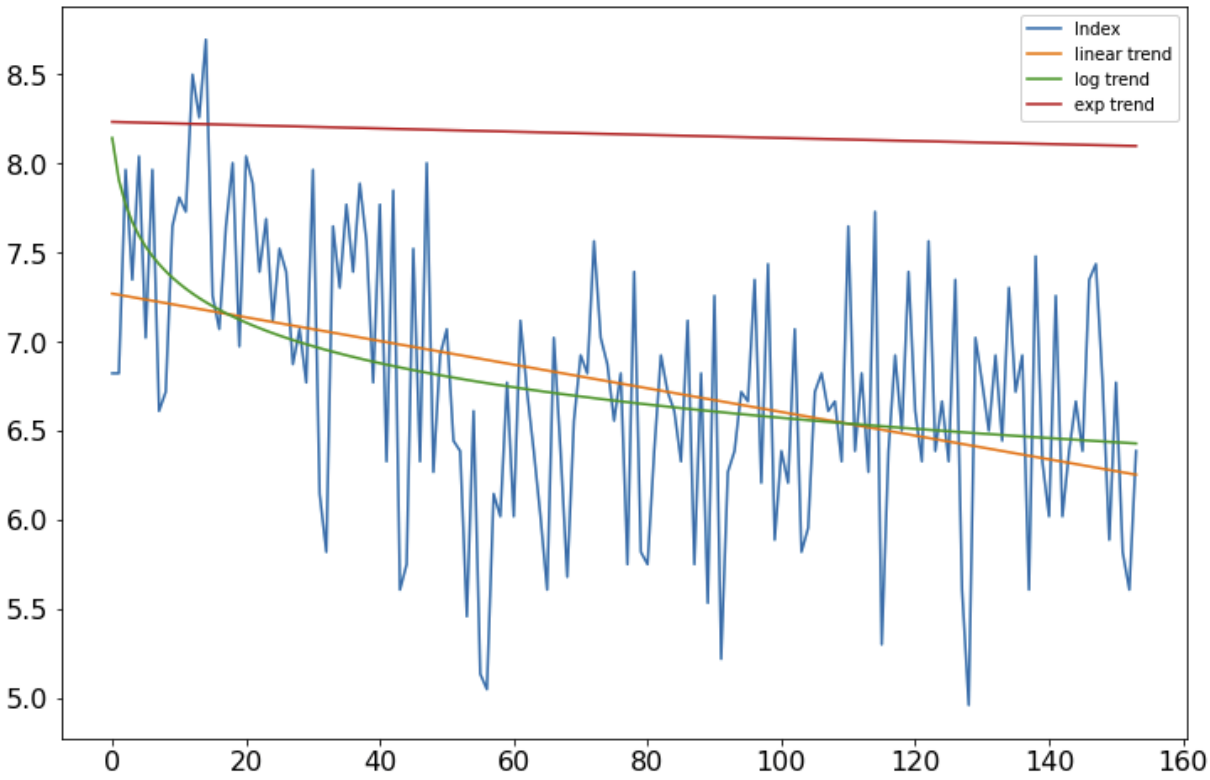
From the print output we can see the best lambda is 0.3542288312735581, which is between a log transform lamda value of 0.0 and a square root transform lambda value of 0.5.

In conclusion, we can see from the legend that the original skew has been reduced from 0.3374 (red) to -0.0227 (green) as well with the box-cox transformation and therefore handling heteroskedasticity.

Removing Trends

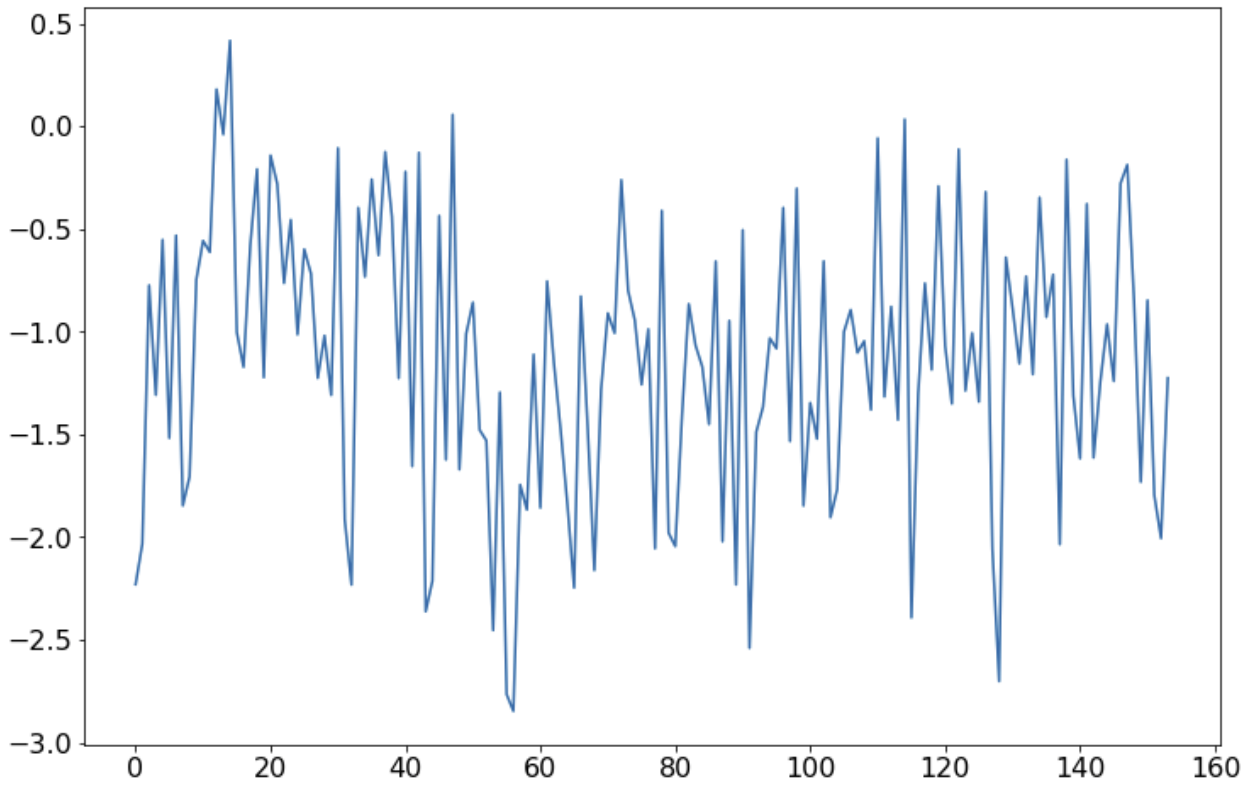
A trend makes a time series non-stationary by increasing the level. This has the effect of varying the mean time series value over time. Lets continue to make the time series stationary by plotting multiple trend lines to find the best fit.

Dealership Trend Analysis

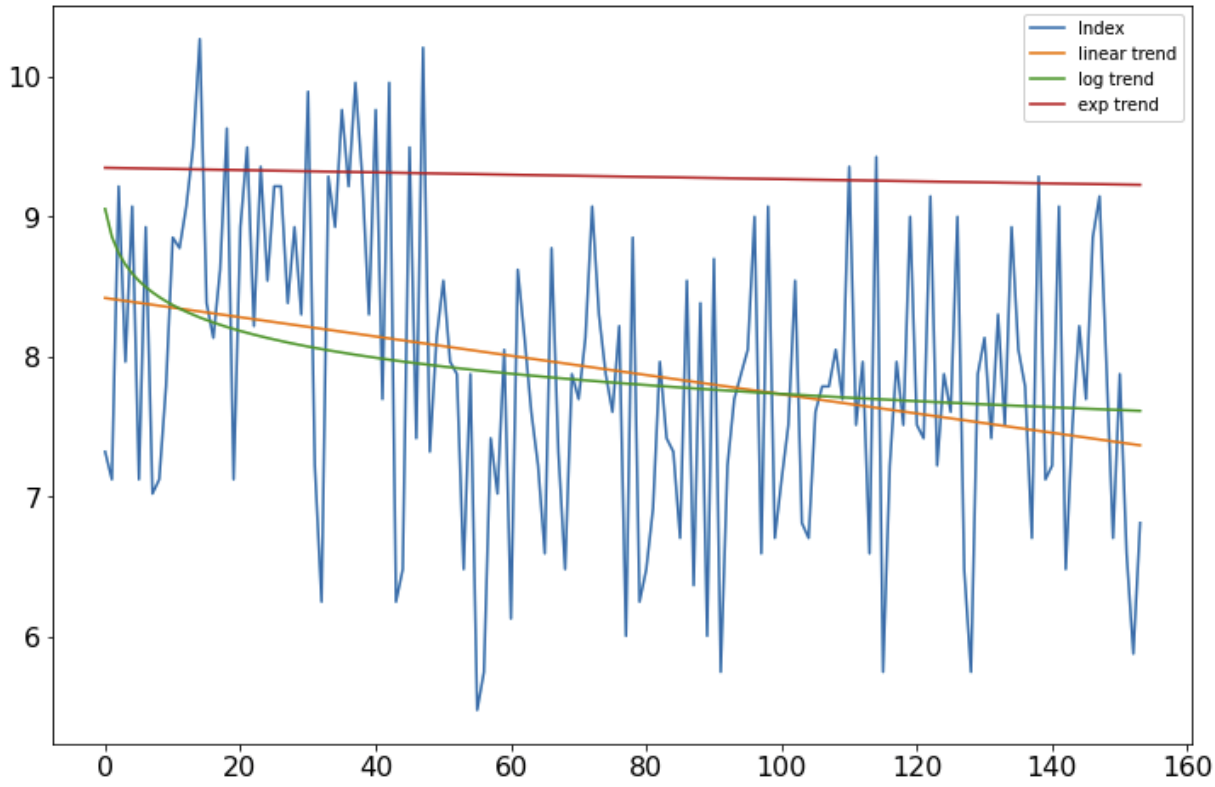


The logarithmic trend is the best fit in this case. Therefore, our new time series becomes: (bcx_dlrshp - logarithmic trend).

(Removal of Logarithmic Trend)

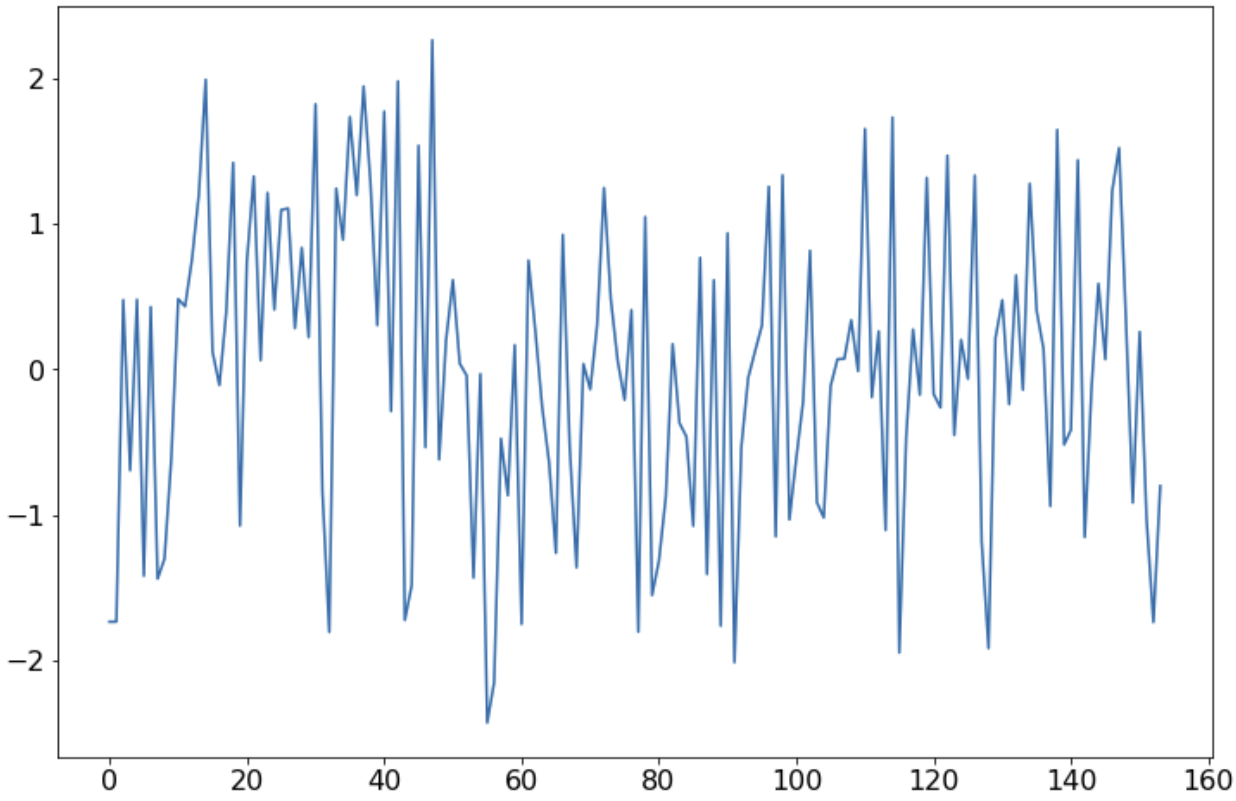


Lexus Trend Analysis



The logarithmic trend is the best fit in this case. Therefore, our new time series becomes: (bcx_lexus - logarithmic trend). Now, let's apply this insight to new time series.

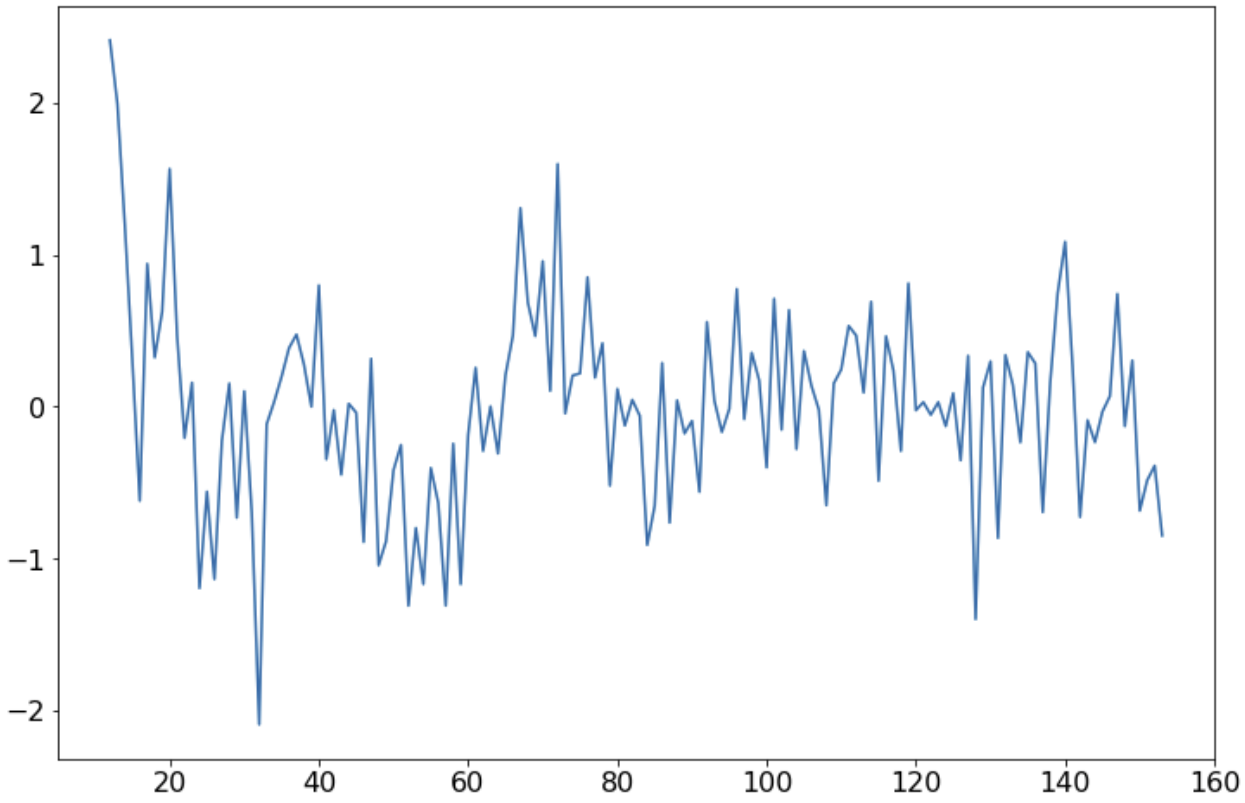
(Removal of Logarithmic Trend)


2
1
0
−1
−2
0
20
40
60
80
100
120
140
160

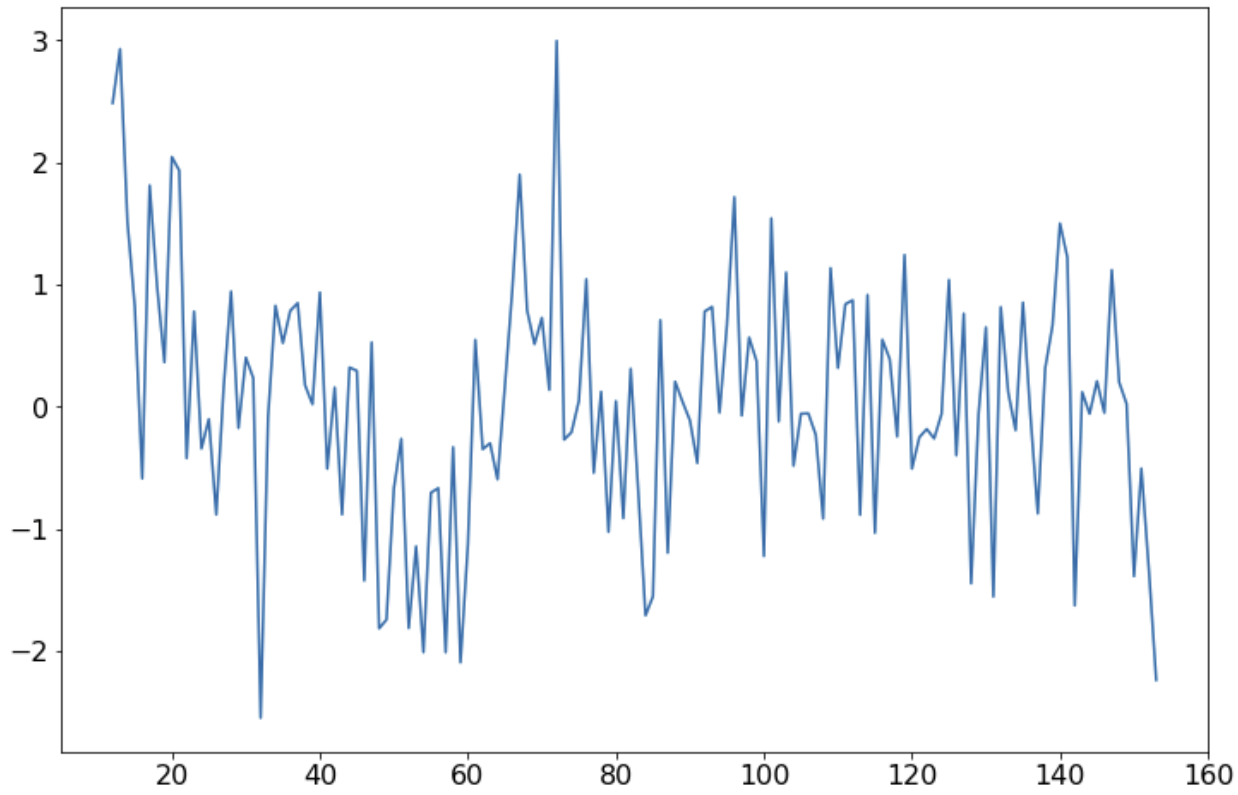
5. Removing Seasonality

Seasonality in time series denotes a recurrent pattern over time. When a series is seasonal, it means that the value at a given point in the past is really close to the value we observe today. In the graph above, it seems to be the case.

Dealership Seasonality



Lexus Seasonality



We have removed most of the trend here, and remain with a stationary series. To make sure that our series is stationary, we can look at the plot: There seems to be no recurrent pattern in the data, constant variance and mean, no trend. So, let's compute the ADFuller test again to confirm.

Dealership Second ADFuller Test

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

Lexus Second ADFuller Test

```
ADF Statistic: -4.149738
p-value: 0.000801
Critical Values:
        1%: -3.483
        5%: -2.884
        10%: -2.579
```

The p-value for both ADF tests are basically 0 and therefore the ADF Statistic is below the 1% critical value for both the dealership and Lexus related datasets. We can now reject the null hypothesis that the series has a unit root and is not stationary. Therefore, these time series are now stationary and ready for modeling.

Summary of Exploratory Data Analysis

- The exploratory data analysis of the Historical Sales Dataset was explored as a whole representing the dealership's total sales and at the brand level, focusing on Lexus as the . number one selling brand for the dealership. Through this layered analysis it is confirmed that the year 2009 had the most statistically significant difference (negatively) in sales performance in comparison to 2004 through 2007. 2017 also has a notable statistically significant difference (negatively) in sales performance in comparison to 2004 through 2006 (who have the highest sales performance, especially 2005). In addition, 2013 and 2014 are the most statistically similar with -0.0033 mean difference. As well as, 2008 and 2015 with -0.0093 mean difference.

- Overall, the Seasonality of the time series is downward which was most clearly observed through a multiplicative model decomposition plot in section 3.5.3. This seasonality displays a clear downward trend yearly beginning in 2006 and a clear increasing fluctuation seasonally throughout the first two quarters of the year for the sale of all dealership vehicles. However, there is a clear downward trend yearly beginning in 2008 and a clear increasing fluctuation seasonally throughout the year for the sale of Lexus vehicles.

- During the analysis of sales performance by vehicle make, in section 3.5.4, Lexus is confirmed to statistically be the highest selling vehicle make in this time series. Within the Lexus brand model types "RX 350", "ES 350", and "RX 330" are the top three selling models in that order.

- Finally, four files have been saved from this notebook. The two are the "TotalSales" and "LexusSales" which are saved in csv format from the "df_total_sales" dataframe created in section 3.5.1.1. These datasets will be used in the next notebook to train and execute various forecasting models. The remaining two files have been saved as two separate time series objects. The first contains the total sales for the dealership, titled "dlrshp_stationary", and the second contains the total sales for Lexus vehicles only, titled "lexus_stationary". Both are in pickle format for panadas objects. Since, "dlrshp_stationary" and "lexus_stationary" were transformed step-by-step into a stationary time series within this notebook, they will be modeled and used as a comparison of performance against models that applied the non-stationary data from file "df_total_sales" in the next notebook.

---

6. Preprocessing and Training

In the prior section we examined the "Historical Sales Dataset" and transformed it step-by-step into a stationary time series. This transformation allowed us to visualize the distribution, heteroskedasticity, trend and seasonality of the time series for the dealership at large and Lexus specifically. Now, it is time to use this insight to train and model for the forecasting of future sales.

This section will begin with four simple baseline models and then transition to candidate models, all in the goal of using historical time series data in a statistical model to forecast future vehicle sales by make and model for this population based on past results. This notebook will compare the performance and accuracy of the various models using the mean absolute percentage error (MAPE). In addition, the two step-by-step transformed stationary datasets from the prior notebook will be modeled as well for further comparison of performance and accuracy to the traditional methods used prior.

6.1 Baseline Models

A baseline in forecast performance provides a point of comparison.

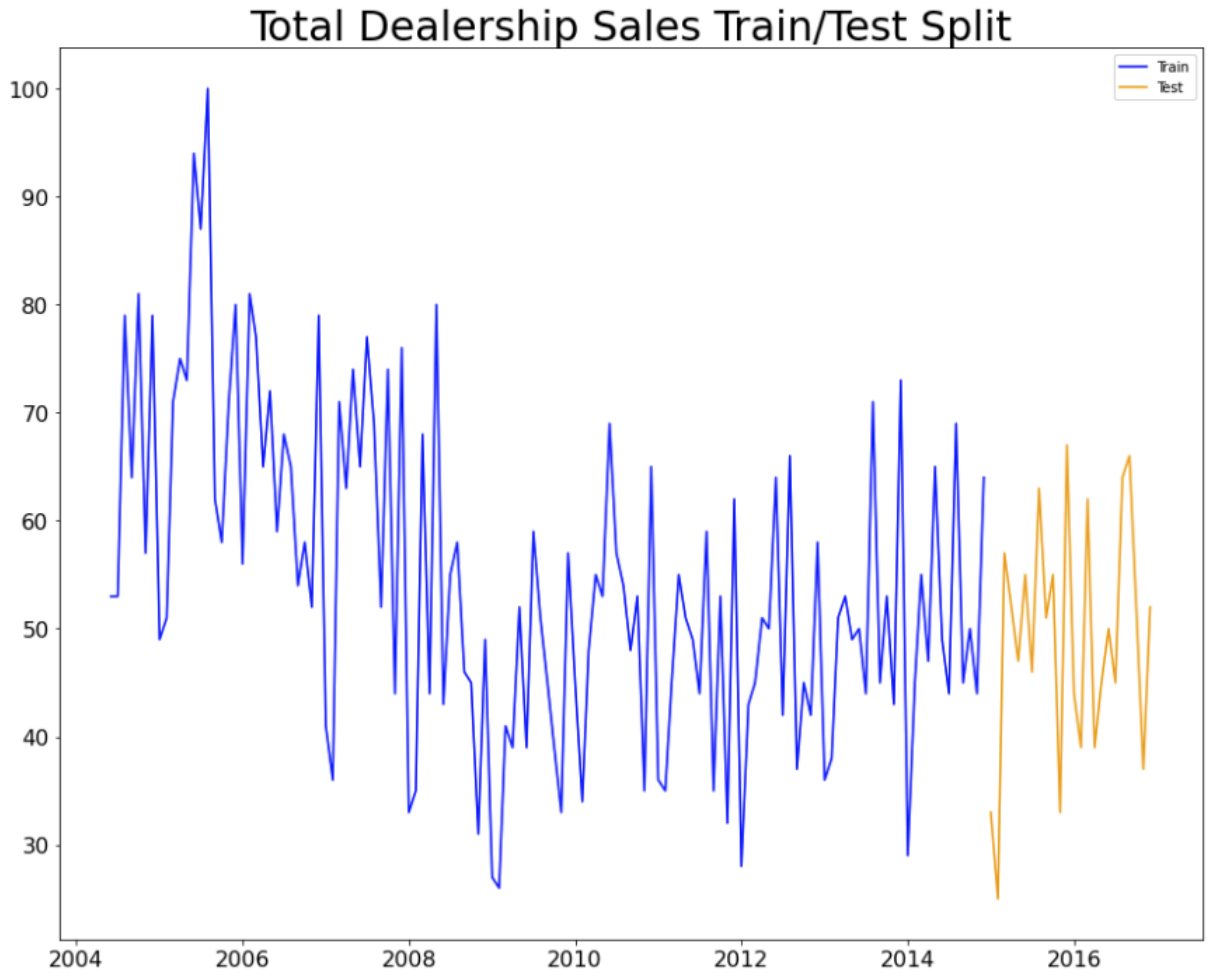
It is a point of reference for all other modeling techniques on our problem. If a model achieves performance at or below the baseline, the technique should be fixed or abandoned.

The goal is to get a baseline performance on our time series forecast problem as quickly as possible so that you can get to work better understanding the dataset and developing more advanced models.

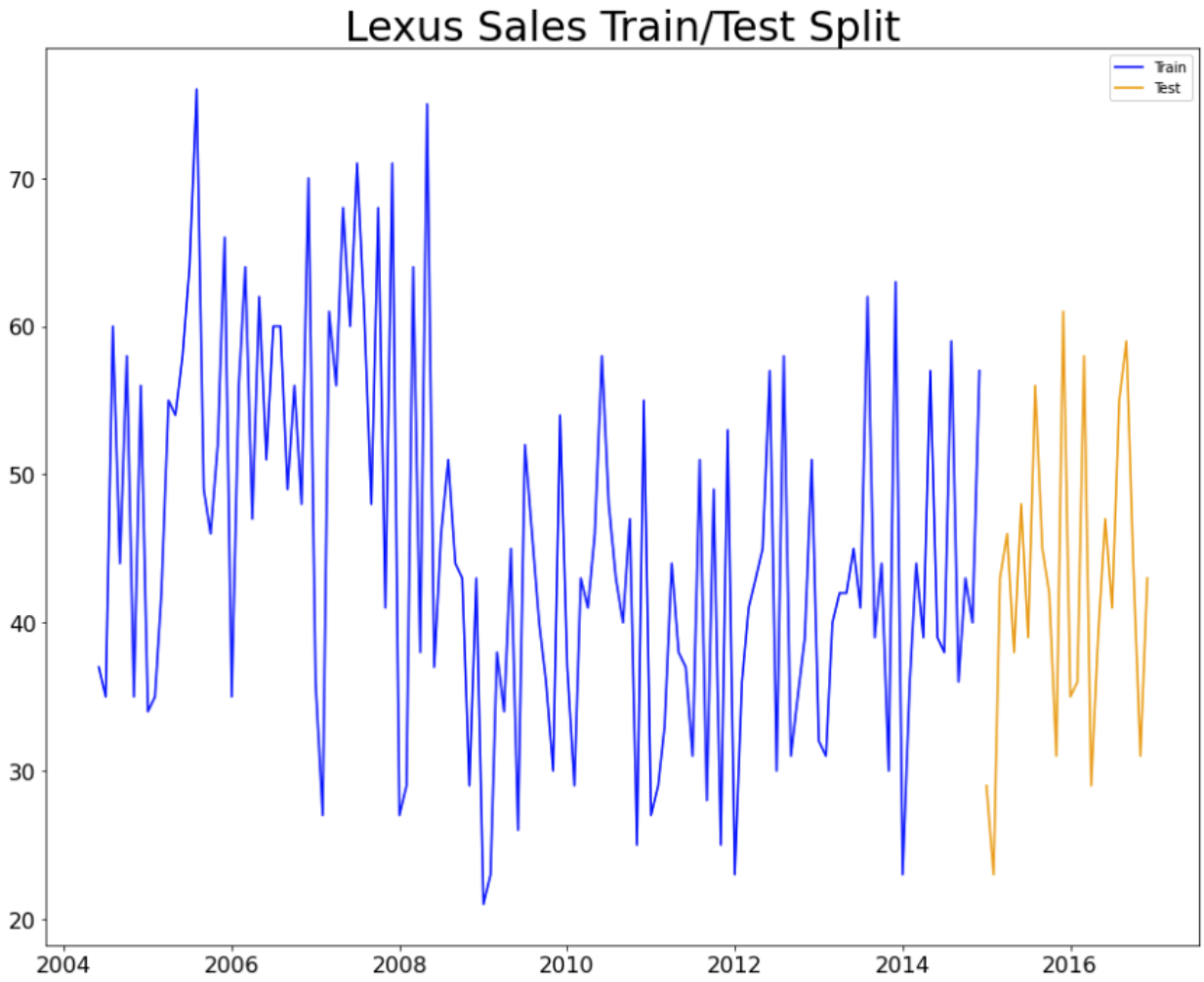
In this section, the following four methods will be explored as baseline models: Simple Average, Naive, SARIMAX, and Holts-Winter's.

Train/Test Split

Dealership Split



Lexus Split

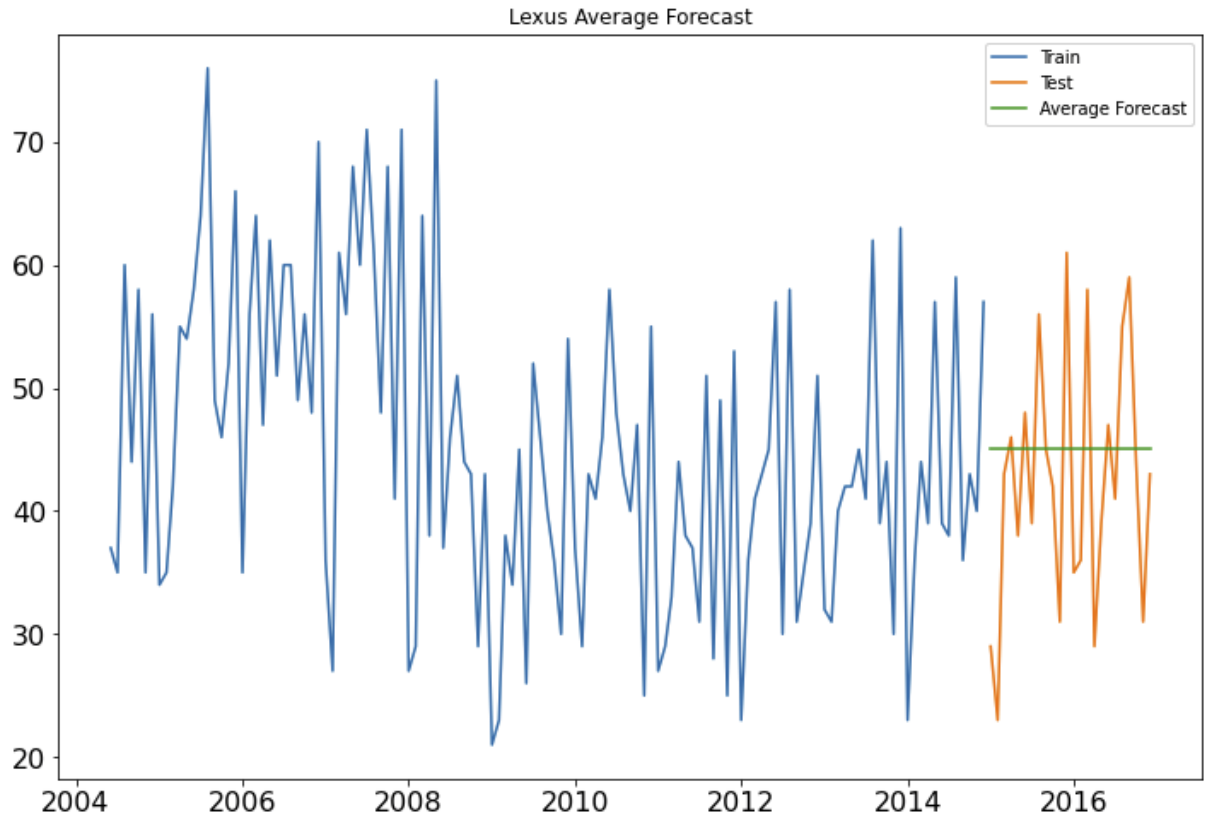


Method 1: Simple Average

With the Simple Average method forecasts of all future values are equal to the average (or "mean") of the historical data.

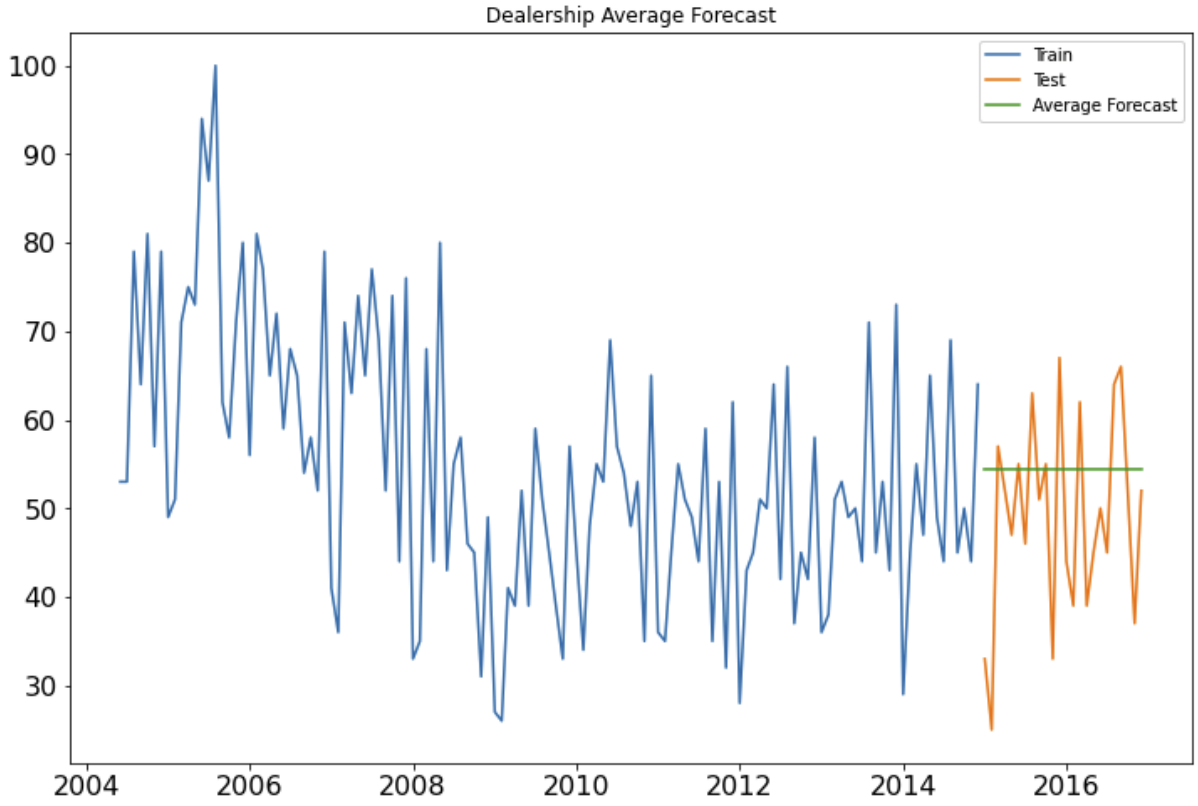
Dealership Average

	Method	RMSE	MAPE
0	Simple Average	10.38	22.993712



Lexus Average

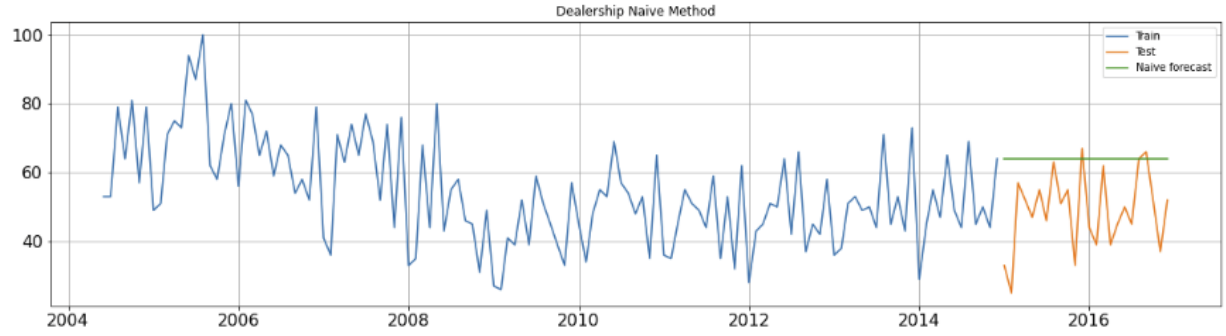
	Method	RMSE	MAPE
0	Simple Average	12.17	24.471189



Method 2: Naive

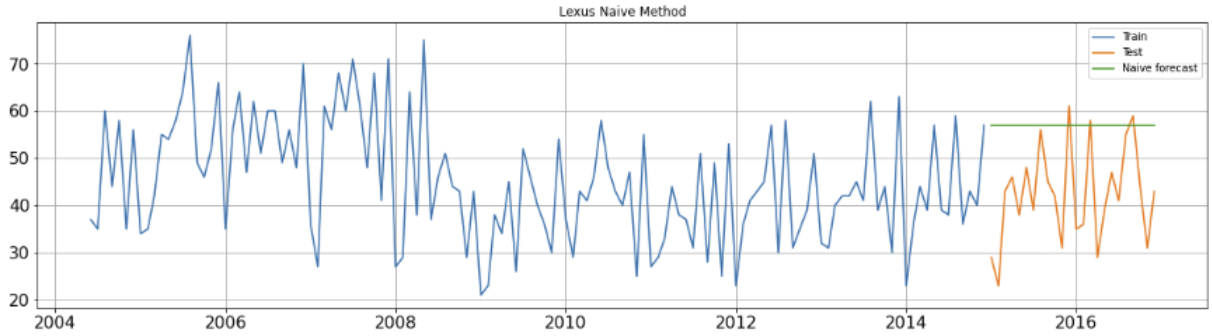
The Naive method is an estimating technique in which the last period's actuals are used as this period's forecast, without adjusting them or attempting to establish causal factors.

Dealership Naive



	Method	RMSE	MAPE
0	Simple Average	12.17	24.471189
0	Naive	18.47	38.762770

Lexus Naive



	Method	RMSE	MAPE
0	Simple Average	10.38	22.993712
0	Naive	17.71	43.719574

Method 3: SARIMAX

Autoregressive Integrated Moving Average, or ARIMA, is one of the most widely used forecasting methods for univariate time series data forecasting.

Although the method can handle data with a trend, it does not support time series with a seasonal component.

An extension to ARIMA that supports the direct modeling of the seasonal component of the series is called Seasonal Autoregressive Integrated Moving Average, or SARIMA.

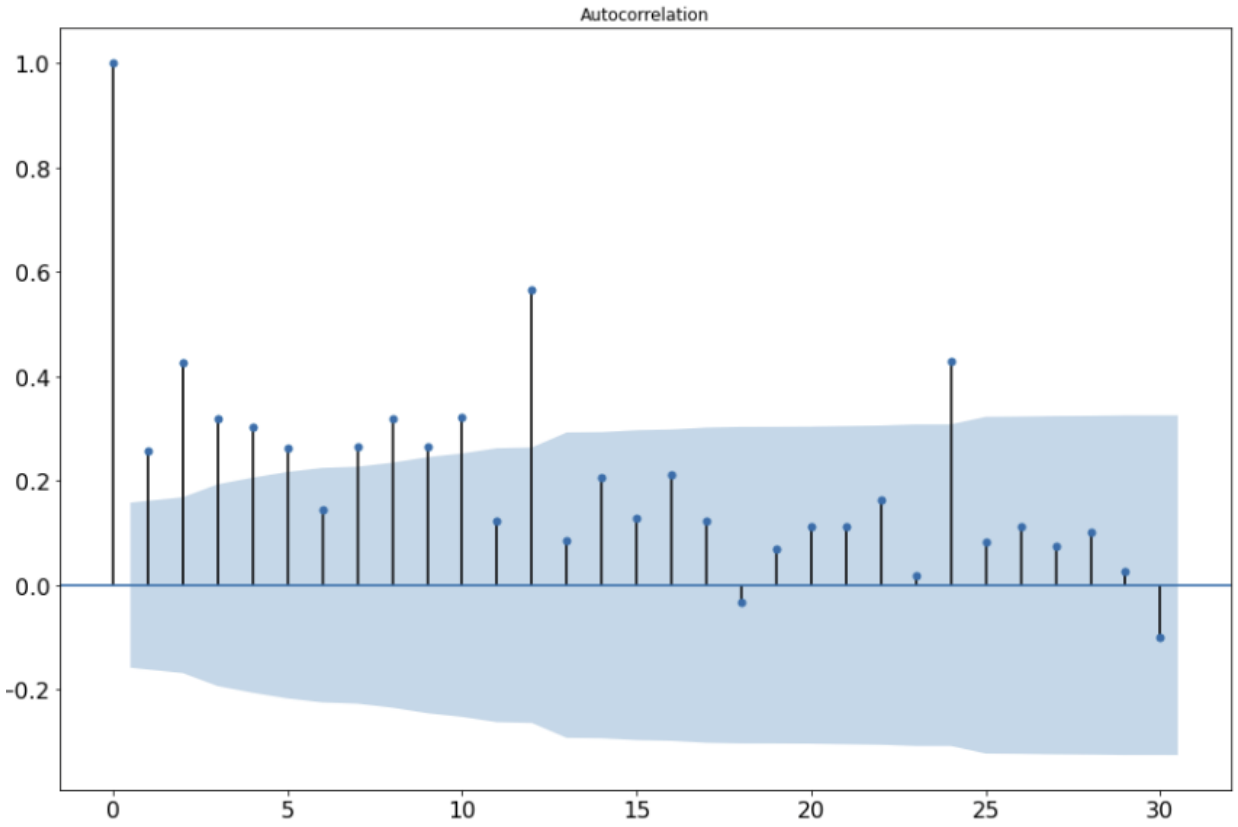
To begin prepping the data for the next model, we must first define one of the key parameters for an ARIMA model, (m). This parameter represents the number of time steps for a single seasonal period and is key in establishing an accurate ARIMA model. To tackle this task, we will use an autocorrelation plot, or ACF.

Autocorrelation (ACF) Plot

Autocorrelation plots are plots that graphically summarize and calculate the correlation for time series observations with observations with previous time steps, called lags. A plot of the autocorrelation of a time series by lag is called the AutoCorrelation Function, or the acronym ACF. Using the plot_acf() function will create a plot showing the lag value along the x-axis and the correlation on the y-axis between -1 and 1. Confidence intervals are drawn as a cone. By default, this is set to a 95% confidence interval, suggesting that correlation values outside of this code are very likely a correlation and not a statistical fluke.

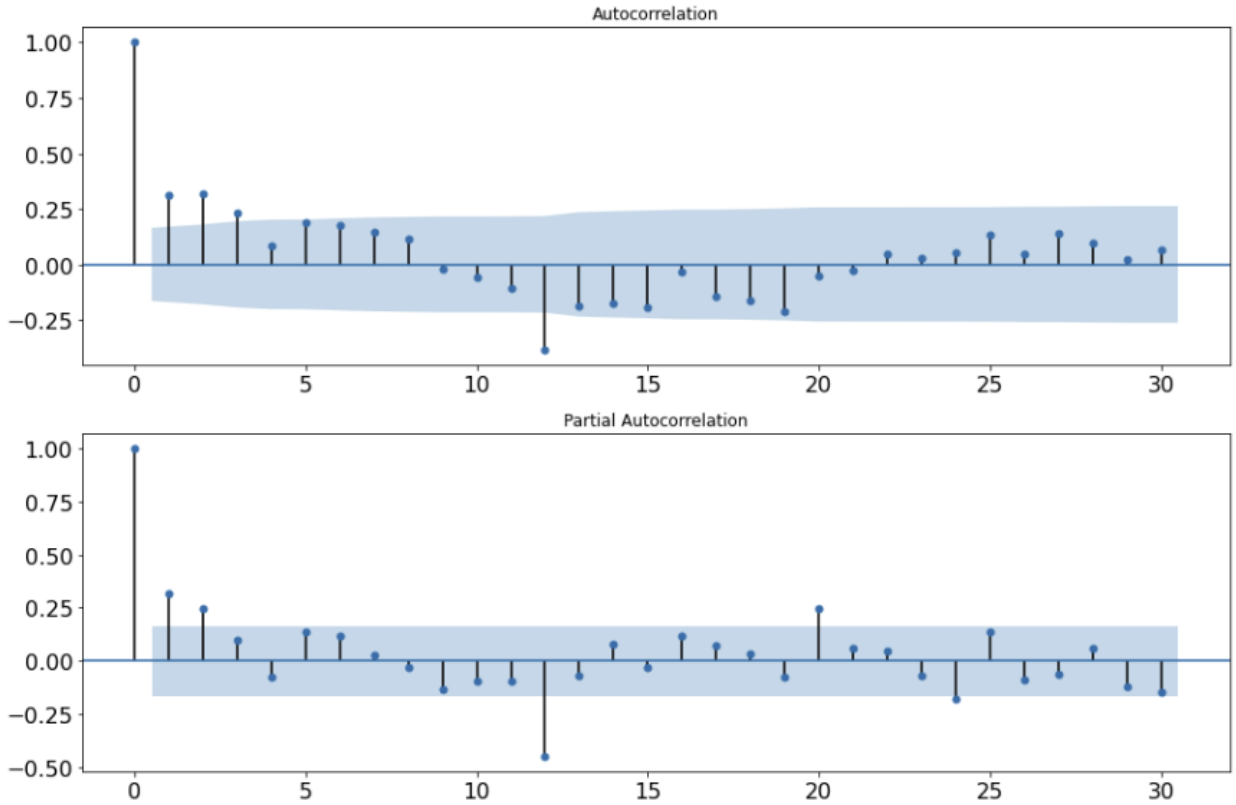
The goal of using the ACF plot is to identify at what lag number we observe the greatest autocorrelation. This lag number will be the value we use for the various terms in our ARIMA model.

Dealership ACF Plot



We can see by the plot above that this time series is not stationary since we see significant correlation at lags 12 and 24. As a result, we will use the stationary version of this dataset generated from the prior notebook and read into this notebook as the "dlr_stationary" dataset. Let's now plot again the ACF and this time also the PACF.

The PACF (Partial Autocorrelation Function) is the correlation between a time series and the lag version of itself after the effect of correlation at smaller lags is subtracted.



To estimate the amount of AR terms we examine the PACF plot above and see lag 1 and 2 are out of the confidence interval. Therefore, we can estimate to use (2) AR terms for our model.

To estimate the amount of I(d) terms is to know how many differencing was used to make the series stationary. In the prior notebook I did used logarithmic to remove trend. Therefore, we can estimate to use (1) I terms.

To estimate the amount of MA terms, this time you will look at ACF plot and see lags 1, 2, and 3 are out of the confidence interval. Therefore, we can estimate to use (3) MA terms for our model.

Now, since we saw the original dealership dataset had seasonality and we need to therefore, use a Seasonal ARIMA model. To do this we need to estimate the seasonal terms for AR, I, and MA as well. To estimate the amount of "seasonal" AR terms, we will look one more time to the PACF function. Now, instead of counting how many lollipops are out of the confidence interval, we will count only how many seasonal lollipops are out. From the PACF plot above we can see (2) seasonal lollipops (12 and 24) are outside of the confidence interval. So, we will add (1) term for seasonal AR (SAR).

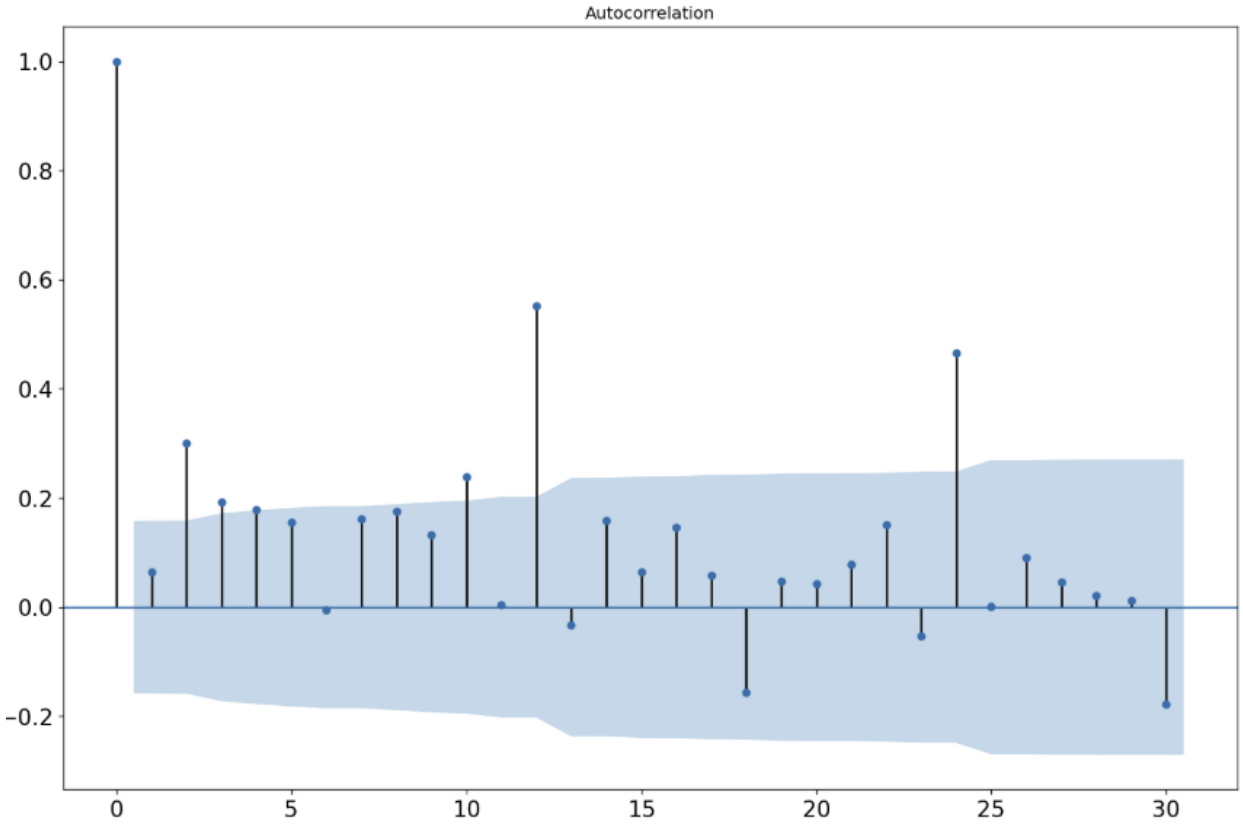
Since seasonal differencing was used to make this time series stationary in the prior notebook. We will estimate (1) seasonal I term for differencing.

To estimate the amount of seasonal MA terms, we will look at the ACF plot and identify significant correlation at lag 12. Therefore, we will estimate (1) term for seasonal MA (SMA).

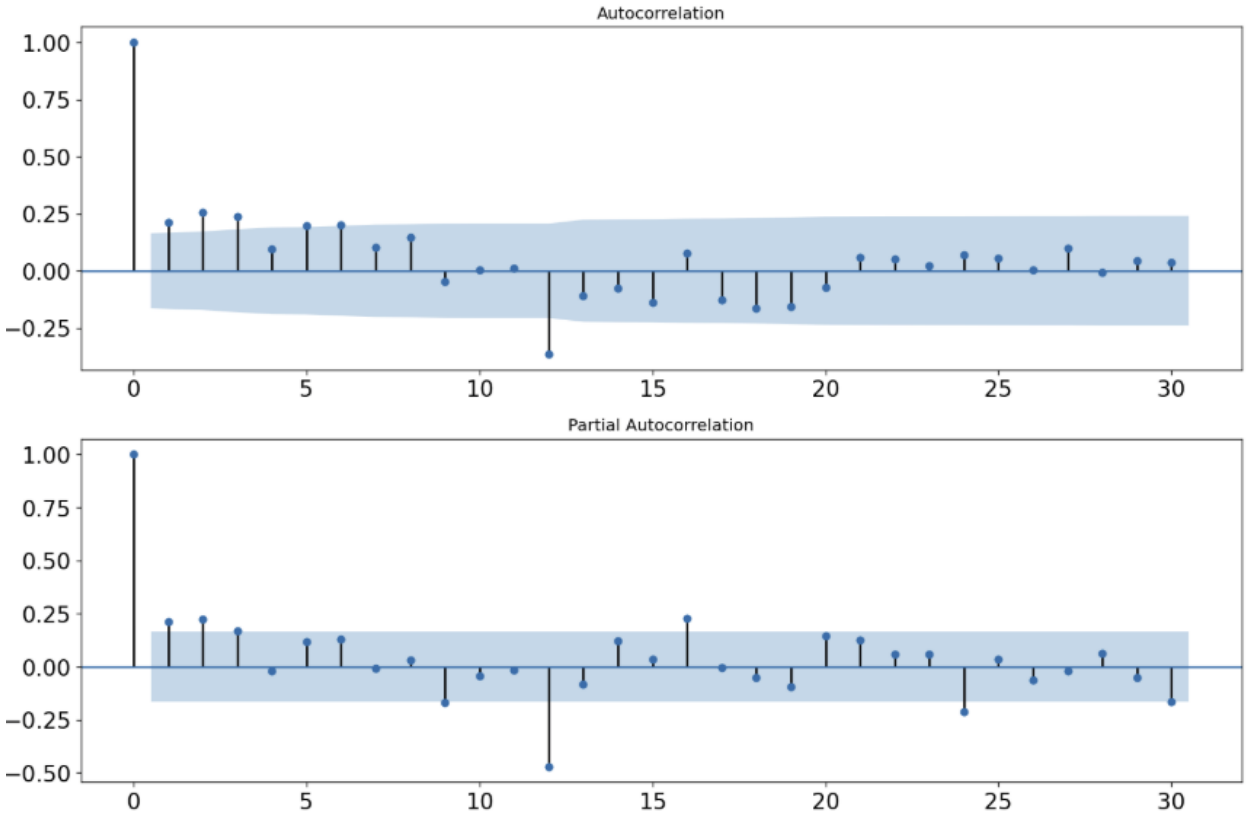
From the ACF plot we can also see that our greatest autocorrelation is at lag number 12. Therefore, we have a (m) value of 12, which makes sense given the data was collected in a monthly basis and we know we have yearly seasonality.

Therefore, we now have all we need to build our SARIMA model. The parameters estimated are (2,1,3),(1,1,1)[12].

Lexus ACF Plot



Again, we can see by the plot above that this time series is not stationary since we see significant correlation at lags 12 and 24. As a result, we will again use the stationary version of this dataset generated from the prior notebook and read into this notebook as the "lex_stationary" dataset. Let's now plot again the ACF and the PACF.



To estimate the amount of each term we examine again either the ACF or PACF plots above. Starting with AR, we examine the PACF and estimate (3) AR terms for our model.

In the prior notebook as stated before, logarithmic was used to remove trend. Therefore, we can estimate to use (1) I terms.

Examining the ACF plot and we can see lags 1, 2, and 3 are again out of the confidence interval. Therefore, we can estimate to use (3) MA terms for our model.

Again, since we saw the original lexus dataset had seasonality we need to therefore, use a Seasonal ARIMA model. From the PACF plot above we can see again (2) seasonal lollipops (12 and 24) are outside of the confidence interval. So, we will add (1) term for seasonal AR (SAR).

Since seasonal differencing was used to make this time series stationary in the prior notebook. We will estimate (1) seasonal I term for differencing again.

To estimate the amount of seasonal MA terms, we will look at the ACF plot and see only one seasonal lag outside of the confidence interval (12). So, therefore we estimate (1) term for seasonal MA (SMA).

From the ACF plot we can also see that our greatest autocorrelation is at lag number 12. Therefore, we have a (m) value of 12.

Now we have what we need to build our SARIMA model. The parameters estimated are (3,1,3),(1,1,1)[12].

Although we have estimated the parameters necessary, we will also use a “grid search” to iteratively explore different combinations of parameters and programmatically select the optimal parameter values for our time series model. Ideally, the grid search will select the same optimal parameters identified using the ACF and PACF plots in this section. However, if not, it will be interesting to compare the model performances using the varying parameters.

6.2 Parameter Estimation (Grid Search)

What is an ARIMA model?

An ARIMA model is a class of statistical models for analyzing and forecasting time series data. It explicitly caters to a suite of standard structures in time series data, and as such provides a simple yet powerful method for making skillful time series forecasts. The 'AR' stands for "autoregression, the 'I' for "integrated, and the 'MA' for "moving average. Each of these components are explicitly specified in the model as a parameter. The parameters of the ARIMA model are defined as follows: (p) is the number of autoregressive lags, (d) is the order of differencing required to make the series stationary, and (q) is the number of moving average lags.

However, we know this data has seasonality as well, so the Seasonal ARIMA model [SARIMAX] will be used to account for seasonal differencing. Therefore, the model we will use is represented by SARIMAX(p,d,q)(P,D,Q)m, where(P,D,Q)m represents the seasonal parameter of (p,d,q) and (m) the number of time steps for a single seasonal period. Since the time series is observed monthly, (m) will be set to 12, representing each month of a year.

How does an auto_arima( ) function work?

To discover the optimal order for an ARIMA model the auto_arima() function from the pmdarima library is utilized. The auto_arima() function works by conducting differencing tests to determine the order of differencing (d), and then fitting models within ranges of defined possibilities. If the seasonal option is enabled, the auto_arima() function also seeks to identify the optimal (P) and (Q) hyperparameters after conducting the Canova-Hansen to determine the optimal order of seasonal differencing (D). In order to find the best model the auto_arima() function optimizes for a given information_criterion, one of {‘aic’, ‘aicc’, ‘bic’, ‘hqic’, ‘oob’}, in this case we will use AIC, and returns the ARIMA which minimizes that value.

The purpose of using the auto_arima() function is not only to help us determine the order of each parameter for our model, but also it chooses the parameters that help make our time series stationary.

Dealership Grid Search

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

Dep. Variable:	y	**No. Observations:**	154
Model:	SARIMAX(2, 1, 1)x(3, 0, 1, 12)	**Log Likelihood**	-565.886
Date:	Sun, 27 Jun 2021	**AIC**	1149.771
Time:	14:45:32	**BIC**	1177.045
Sample:	0	**HQIC**	1160.850
	- 154		
Covariance Type:	opg		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
intercept	-0.0008	0.004	-0.240	0.811	-0.008	0.006
ar.L1	0.1315	0.086	1.537	0.124	-0.036	0.299
ar.L2	0.3016	0.083	3.620	0.000	0.138	0.465
ma.L1	-0.9490	0.032	-29.553	0.000	-1.012	-0.886
ar.S.L12	1.0454	0.222	4.710	0.000	0.610	1.480
ar.S.L24	0.0013	0.133	0.009	0.992	-0.260	0.262
ar.S.L36	-0.0584	0.129	-0.454	0.650	-0.311	0.194
ma.S.L12	-0.8632	0.275	-3.139	0.002	-1.402	-0.324
sigma2	86.3713	13.015	6.636	0.000	60.862	111.880

Ljung-Box (Q):	36.71	**Jarque-Bera (JB):**	0.26
Prob(Q):	0.62	**Prob(JB):**	0.88
Heteroskedasticity (H):	0.43	**Skew:**	0.10
Prob(H) (two-sided):	0.00	**Kurtosis:**	2.96

The output of our grid search suggests that the best ARIMA model is a SARIMAX model with the ordered parameters of (2,1,1)(3,0,1)[12] which yields the lowest AIC value of 1149.77. We are not focused on the BIC value since this is a predictive model (AIC is a better criterion) versus an explanatory model (where BIC would be a better criterion). We also see the Prob(Q) and Prob(JB) values are both greater than 0.05, so we can accept the

assumptions that the residuals are uncorrelated (Prob(Q)) and Gaussian normally distributed (Prob(JB)). We should therefore now consider this to be optimal model parameters out of all the models considered.

Lexus Grid Search

```
Best model:  ARIMA(0,1,1)(1,0,5)[12] intercept
Total fit time: 115.485 seconds
```

Dep. Variable:	y	**No. Observations:**	154
Model:	SARIMAX(0, 1, 1)x(1, 0, [1, 2, 3, 4, 5], 12)	**Log Likelihood**	-554.323
Date:	Sun, 27 Jun 2021	**AIC**	1126.647
Time:	14:48:05	**BIC**	1153.921
Sample:	0	**HQIC**	1137.726
	- 154		
Covariance Type:	opg		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
intercept	-0.0007	0.008	-0.082	0.935	-0.017	0.016
ma.L1	-0.8112	0.048	-16.931	0.000	-0.905	-0.717
ar.S.L12	0.9794	0.060	16.218	0.000	0.861	1.098
ma.S.L12	-0.8088	0.173	-4.666	0.000	-1.148	-0.469
ma.S.L24	0.0178	0.115	0.155	0.877	-0.207	0.242
ma.S.L36	-0.1157	0.117	-0.990	0.322	-0.345	0.113
ma.S.L48	0.0469	0.149	0.315	0.753	-0.245	0.339
ma.S.L60	0.0491	0.144	0.342	0.733	-0.233	0.331
sigma2	73.2490	11.192	6.545	0.000	51.314	95.184

Ljung-Box (Q):	37.93	**Jarque-Bera (JB):**	1.62
Prob(Q):	0.56	**Prob(JB):**	0.44
Heteroskedasticity (H):	0.51	**Skew:**	0.05
Prob(H) (two-sided):	0.02	**Kurtosis:**	2.51

The output of our grid search suggests that the best ARIMA model is a SARIMAX model with the ordered parameters of (0,1,1)(1,0,5)[12] which yields the lowest AIC value of

1126.6. We are not focused on the BIC value since this is a predictive model (AIC is a better criterion) versus an explanatory model (where BIC would be a better criterion). We also see the Prob(Q) and Prob(JB) values are both greater than 0.05, so we can accept the assumptions that the residuals are uncorrelated(Prob(Q))and Gaussian normally distributed(Prob(JB)). We should therefore consider this to be optimal model parameters out of all the models considered.

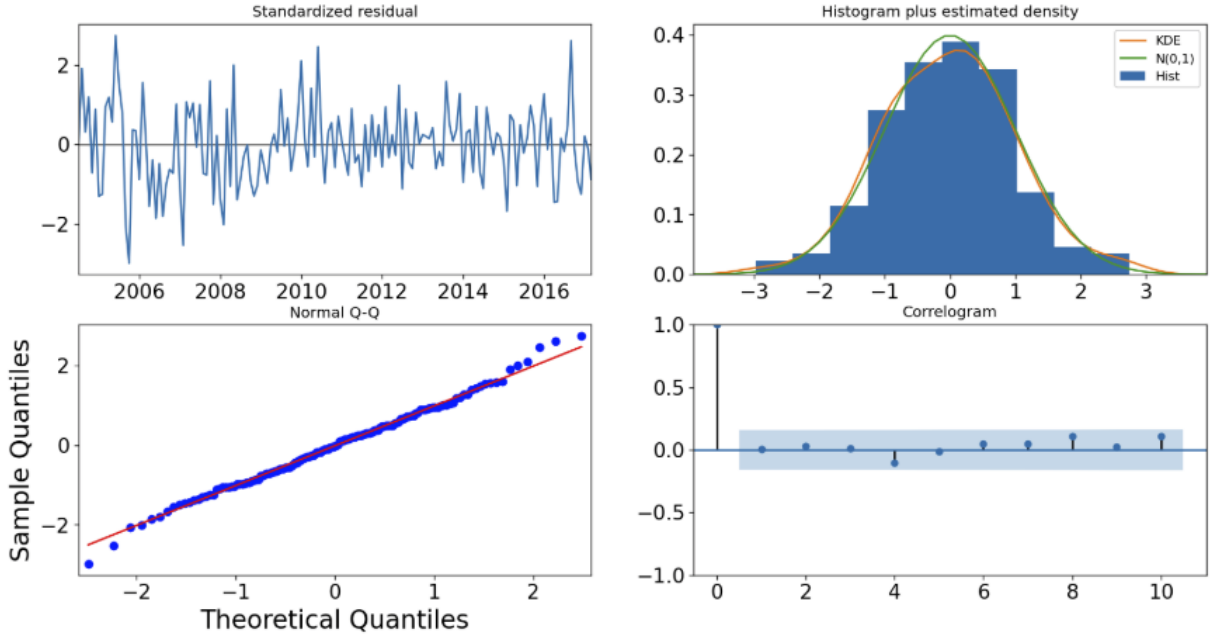
SARIMAX Model Fit

Using grid search, we have identified the set of parameters that produces the best fitting model to our time series data. We can proceed to analyze this particular model in more depth.

We'll start by plugging the optimal parameter values into a new SARIMAX model. Then we will do so again, but instead using the parameters identified while using the ACF and PACF plots.

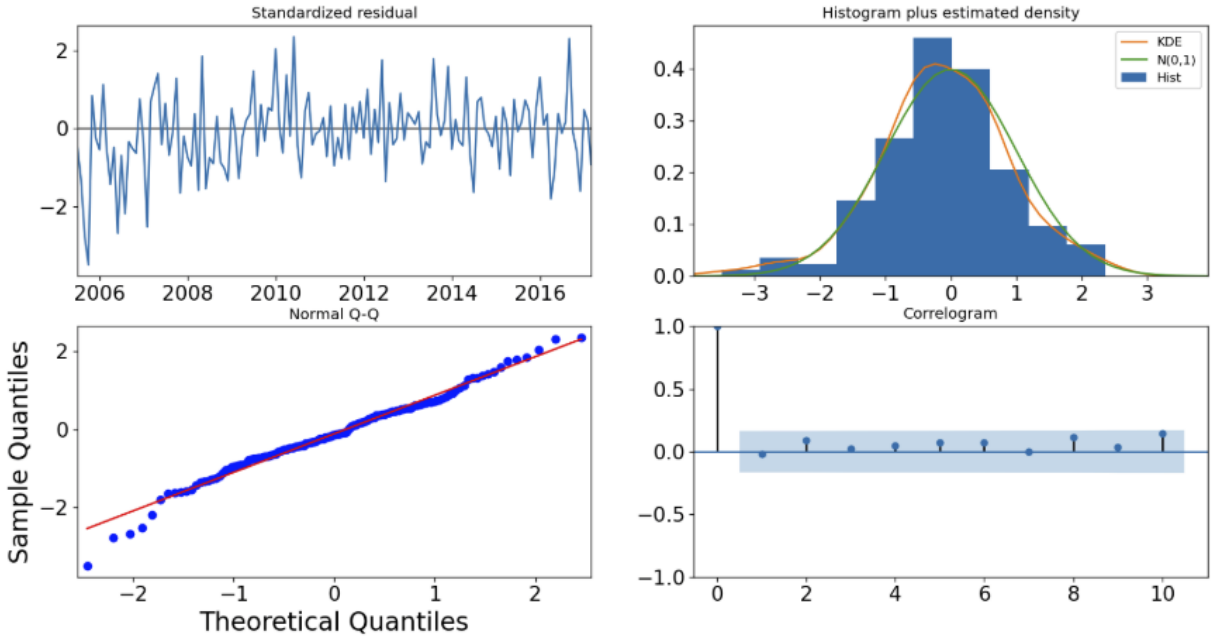
Dealership Model Fit (Grid Search Parameters)

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.0396	0.125	-0.317	0.751	-0.284	0.205
ar.L2	0.1552	0.128	1.213	0.225	-0.095	0.406
ma.L1	-0.8184	0.082	-10.015	0.000	-0.979	-0.658
ar.S.L12	1.0288	0.122	8.463	0.000	0.791	1.267
ar.S.L24	0.0627	0.095	0.663	0.507	-0.123	0.248
ar.S.L36	-0.0924	0.040	-2.332	0.020	-0.170	-0.015
ma.S.L12	-0.9609	0.507	-1.897	0.058	-1.954	0.032
sigma2	79.3088	31.331	2.531	0.011	17.902	140.716



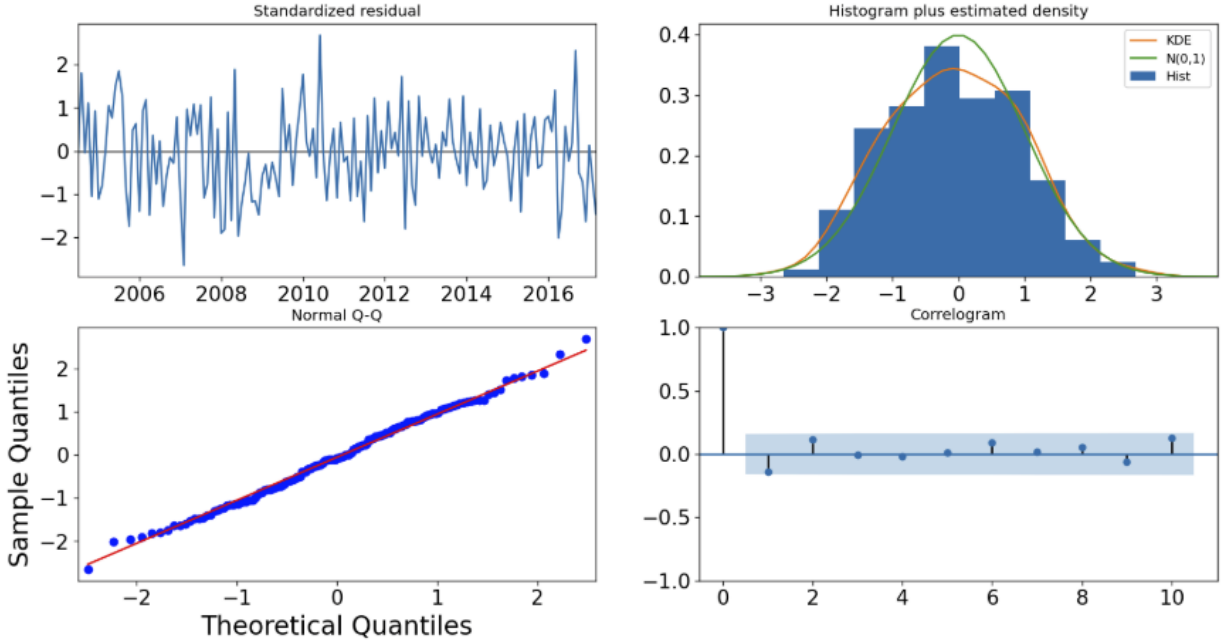
Dealership Model Fit (ACF & PACF Parameters)

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.3487	0.088	3.945	0.000	0.175	0.522
ar.L2	-0.7737	0.092	-8.371	0.000	-0.955	-0.593
ma.L1	-1.1571	0.088	-13.200	0.000	-1.329	-0.985
ma.L2	1.2804	0.203	6.321	0.000	0.883	1.677
ma.L3	-0.7821	0.148	-5.297	0.000	-1.072	-0.493
ar.S.L12	0.0663	0.143	0.462	0.644	-0.215	0.347
ma.S.L12	-0.9777	1.076	-0.908	0.364	-3.087	1.132
sigma2	72.3860	70.254	1.030	0.303	-65.310	210.082



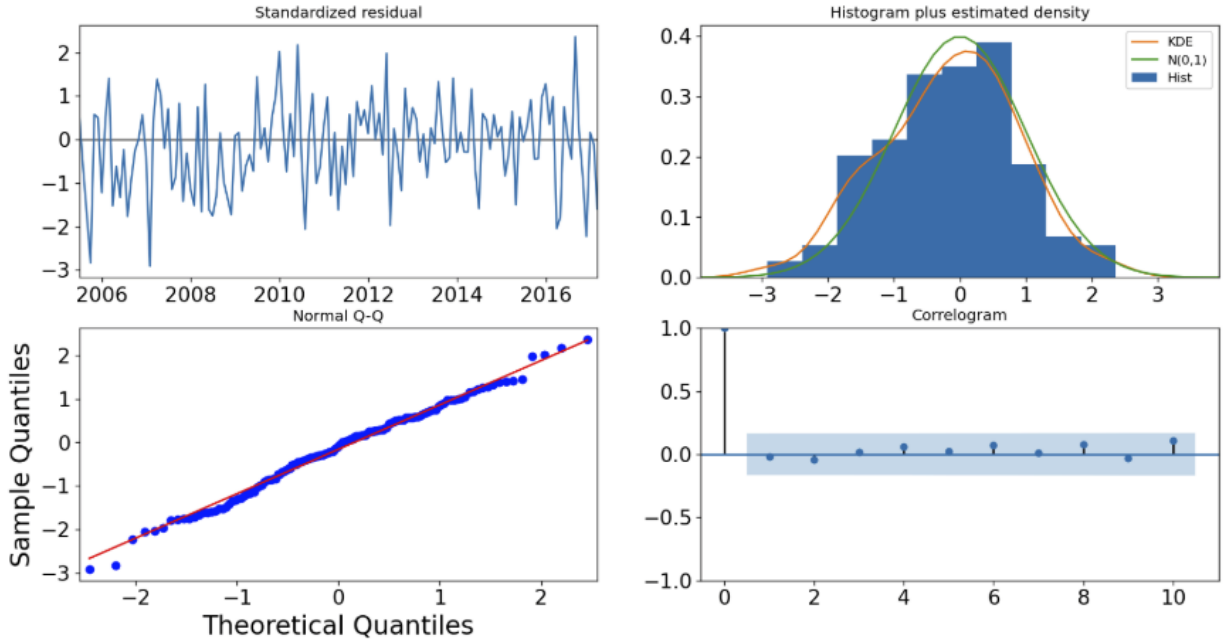
Lexus Model Fit (Grid Search Parameters)

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.8221	0.046	-17.869	0.000	-0.912	-0.732
ar.S.L12	0.9987	0.031	32.520	0.000	0.939	1.059
ma.S.L12	-0.8809	0.481	-1.832	0.067	-1.823	0.061
ma.S.L24	0.0412	0.117	0.353	0.724	-0.188	0.270
ma.S.L36	-0.1477	0.130	-1.134	0.257	-0.403	0.108
ma.S.L48	0.0102	0.146	0.070	0.944	-0.276	0.297
ma.S.L60	0.0288	0.123	0.234	0.815	-0.213	0.271
sigma2	68.9731	27.807	2.480	0.013	14.473	123.474



Lexus Model Fit (ACF & PACF Parameters)

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.1885	0.095	1.985	0.047	0.002	0.375
ar.L2	-0.7011	0.085	-8.224	0.000	-0.868	-0.534
ar.L3	-0.2297	0.104	-2.205	0.027	-0.434	-0.026
ma.L1	-1.1602	0.080	-14.521	0.000	-1.317	-1.004
ma.L2	1.2756	0.086	14.806	0.000	1.107	1.444
ma.L3	-0.7314	0.087	-8.376	0.000	-0.903	-0.560
ar.S.L12	-0.0028	0.126	-0.022	0.982	-0.249	0.243
ma.S.L12	-0.9456	0.359	-2.630	0.009	-1.650	-0.241
sigma2	57.5414	17.699	3.251	0.001	22.852	92.231



From the summary tables for both datasets we can see that the coef column shows the weight (importance) of each feature and how each one impacts the time series. The P>|z| column informs us of the significance of each feature weight. Here, most of the weights have a p-value lower than 0.05, so it is reasonable to retain most of them in our model.

Our primary concern now, is to ensure that the residuals of our model are uncorrelated and normally distributed with zero-mean. If the seasonal ARIMA model does not satisfy these properties, it is a good indication that it can be further improved.

In this case, our model diagnostics suggests that both model residuals are normally distributed based on the following:

- In the top left plot, we see the residuals over time don't display any obvious seasonality and appear to be white noise. This is confirmed by the autocorrelation (correlogram) plot on the bottom right, which shows that the time series residuals have low correlation with lagged versions of itself.

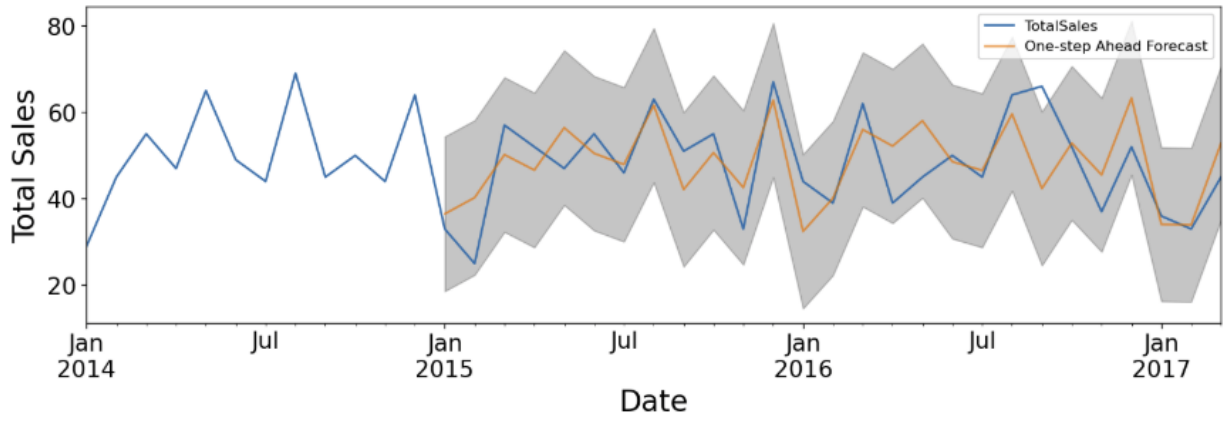
- In the top right plot, we see that the orange KDE line follows closely with the N(0,1) line (where N(0,1)) is the standard notation for a normal distribution with mean 0 and standard deviation of 1). This is a good indication that the residuals are normally distributed.
- The normal qq-plot on the bottom left shows that the ordered distribution of residuals (blue dots) follows the linear trend of the samples taken from a standard normal distribution with N(0, 1). Again, this is a strong indication that the residuals are normally distributed.
- The correlogram shows 95% of the correlations for lag greater than zero do not reach beyond the confidence band and therefore, aren't significant. Which means that the model hasn't failed to capture any of the data.

Those observations lead us to conclude that both of our models produce a satisfactory fit that could help us understand our time series data and forecast future values.

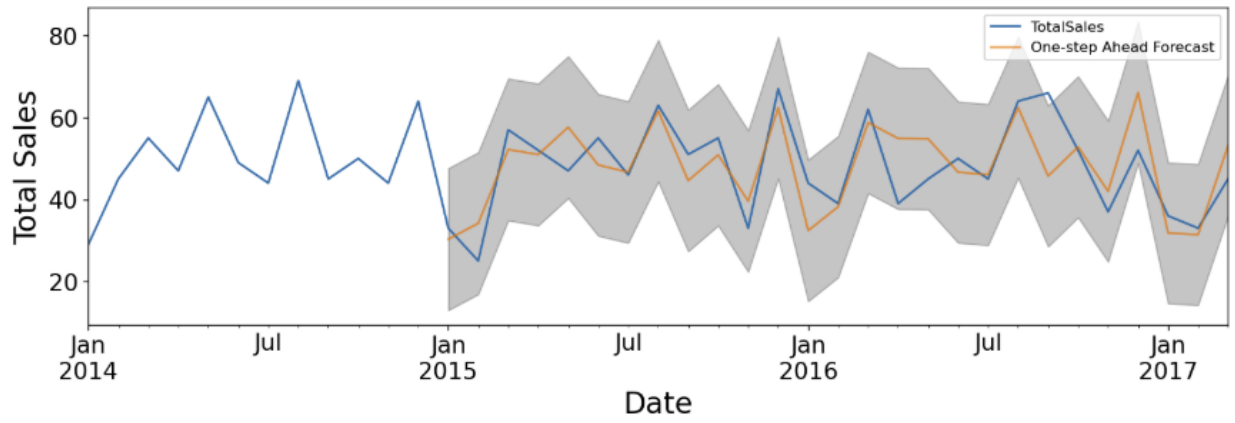
Validating SARIMAX Predictions

We have obtained a model for both of our time series that can now be used to produce forecasts. We start by comparing predicted values to real values of the time series, which will help us understand the accuracy of our predictions. The get_prediction() and conf_int() attributes allow us to obtain the values and associated confidence intervals for forecasts of the time series.

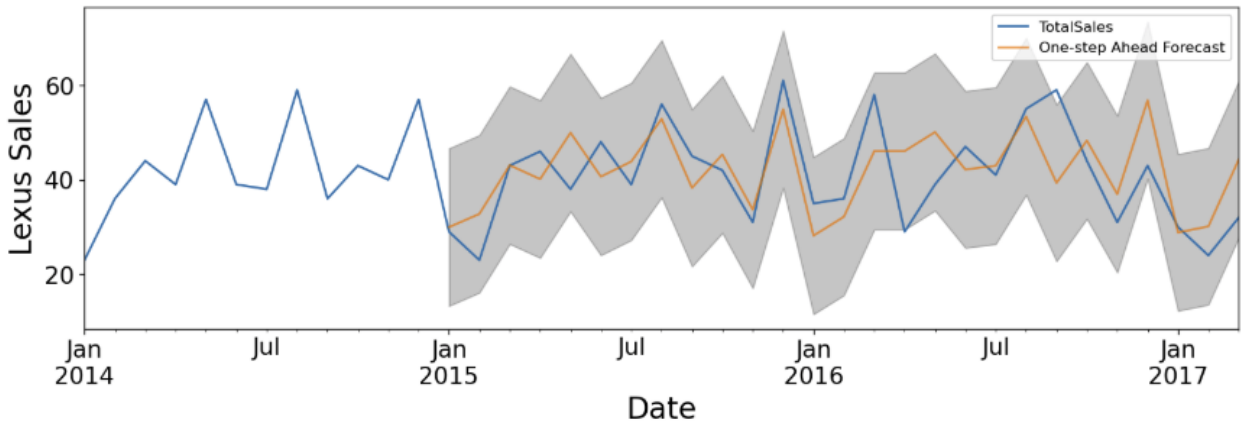
Dealership Prediction (Grid Search Parameters)



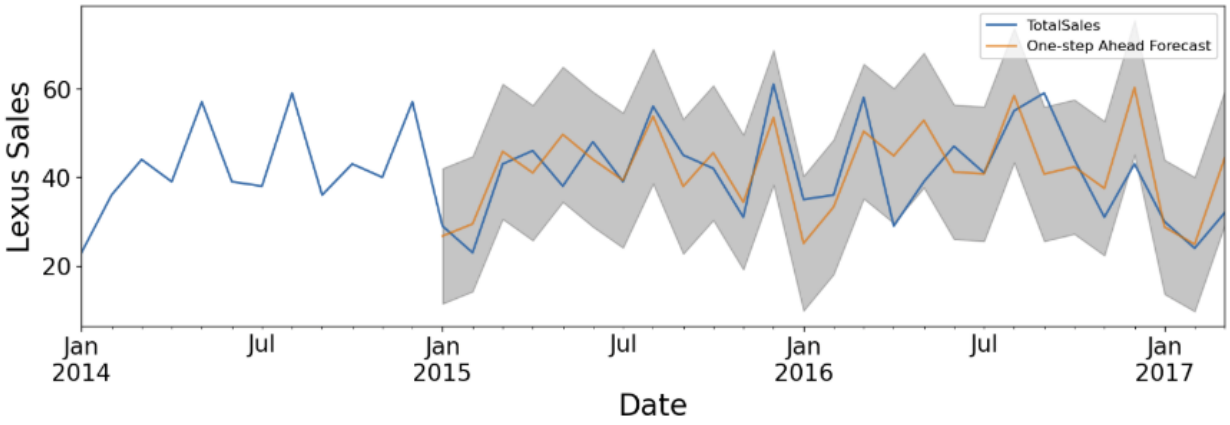
Dealership Prediction (ACF & PACF Parameters)



Lexus Prediction (Grid Search Parameters)



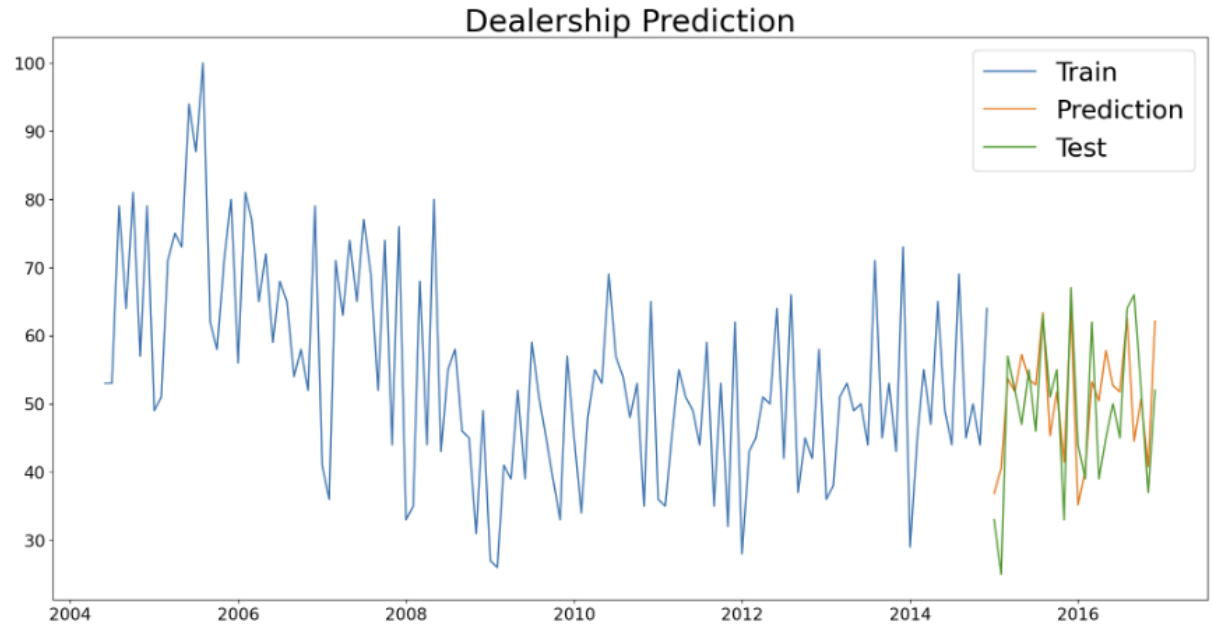
Lexus Prediction (ACF & PACF Parameters)



This step compared the true values with the forecasted predictions. Our predictions fit with the true values fairly well. The command "model_fit.get_prediction(start=pd.to_datetime('2015-01'), dynamic=False)" determined the period to predict within compared with the true data from that same period. Now, let's train/test split our data on our defined model and then forecast future sales.

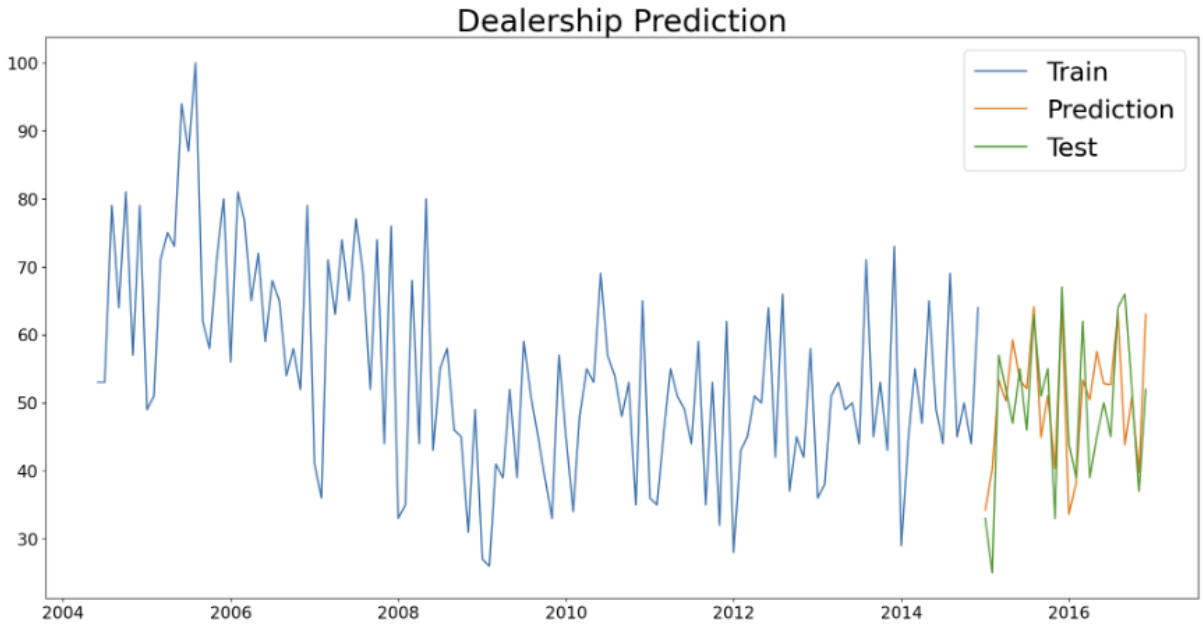
SARIMAX Model Training

Dealership Training (Grid Search Parameters)



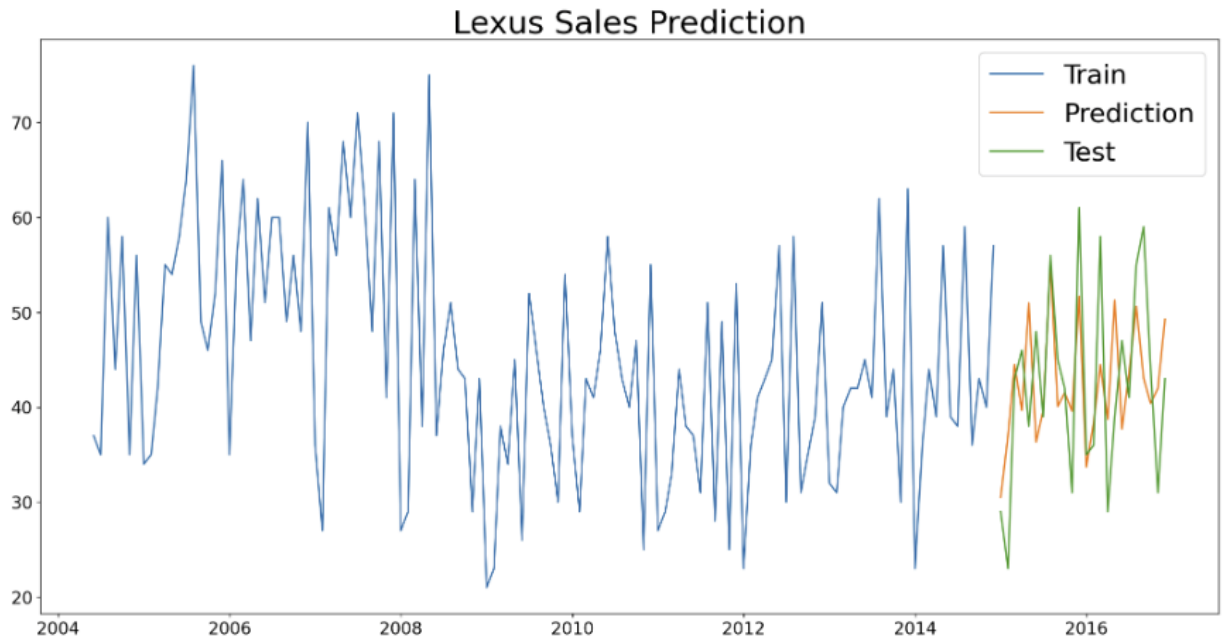
	Method	RMSE	MAPE
0	Simple Average	12.170000	24.471189
0	Naive	18.470000	38.762770
0	SARIMAX	8.241263	14.562545

Dealership Training (ACF & PACF Parameters)



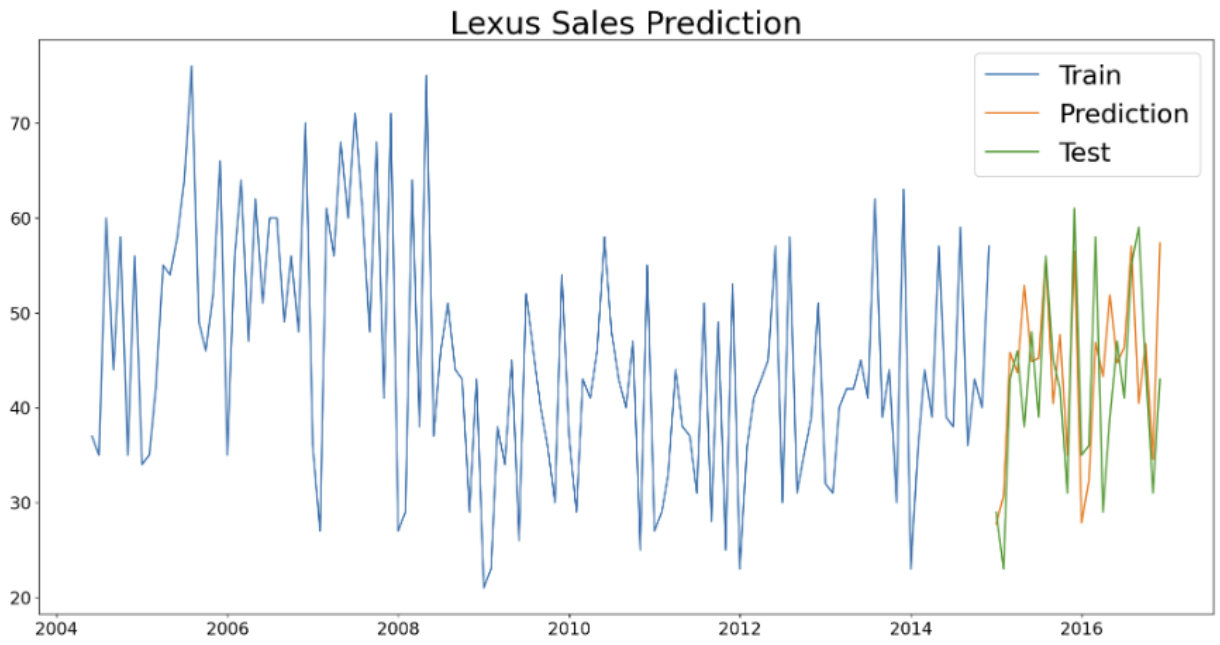
	Method	RMSE	MAPE
0	Simple Average	12.170000	24.471189
0	Naive	18.470000	38.762770
0	SARIMAX	8.241263	14.562545
0	SARIMAX_autoc	8.455708	14.561804

Lexus Training (Grid Search Parameters)



	Method	RMSE	MAPE
0	Simple Average	10.380000	22.993712
0	Naive	17.710000	43.719574
0	SARIMAX	8.428568	17.403573

Lexus Training (ACF & PACF Parameters)



	Method	RMSE	MAPE
0	Simple Average	10.380000	22.993712
0	Naive	17.710000	43.719574
0	SARIMAX	8.428568	17.403573
0	SARIMAX_autoc	8.161580	16.316625

Method 4: Holt-Winter's

A Holt-Winter's Method (a.k.a. Triple Exponential Smoothing Model) extends Holt to allow the forecasting of time series data that has both trend and seasonality. A Holt-Winter's Exponential Smoothing Model subsumes single and double exponential smoothing by the configuration of the nature of the trend (additive, multiplicative, or none) and the nature of the seasonality (additive, multiplicative, or none), as well as any dampening of the trend.

In this section, we will develop a framework for grid searching exponential smoothing model hyperparameters for our given univariate time series.

Running the block of code below is relatively slow given the large amount of data. Model configurations and the RMSE will be printed as the models are evaluated. The top three model configurations and their error will be reported at the end of the run.

Parameter Estimation (Grid Search)

Dealership Grid Search

The best result out of the top three is an RMSE of about 8.220 degrees with the following configuration:

- **Trend:** Multiplicative
- **Damped:** False

- **Seasonal:** Multiplicative
- **Seasonal Periods:** 12
- **Box-Cox Transform:** True
- **Remove Bias:** False

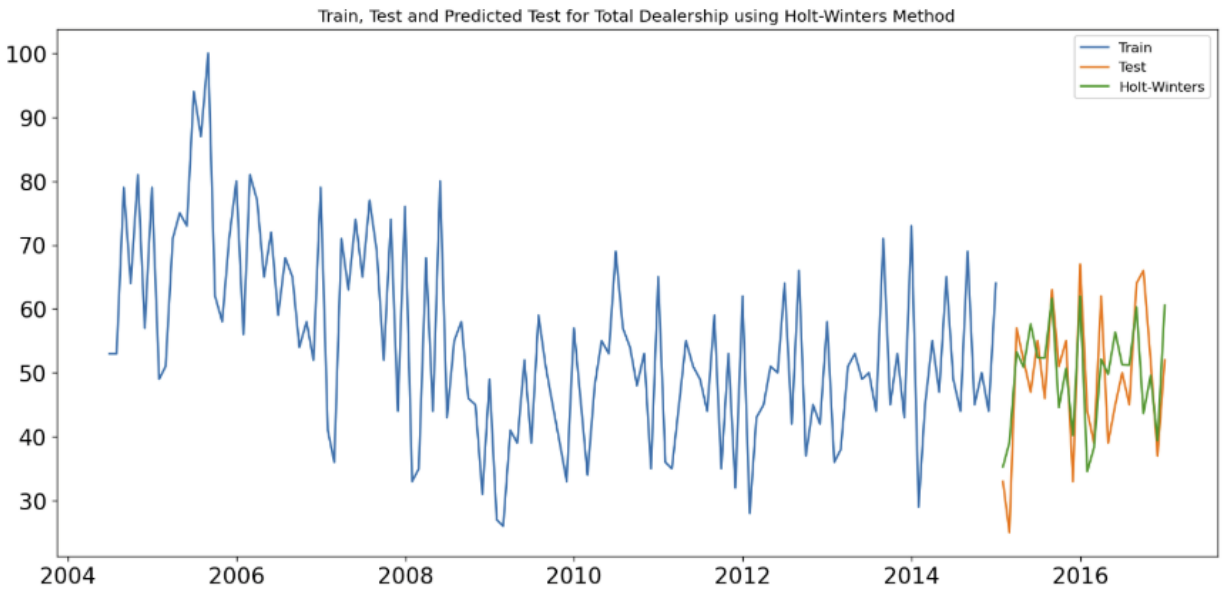
Lexus Grid Search

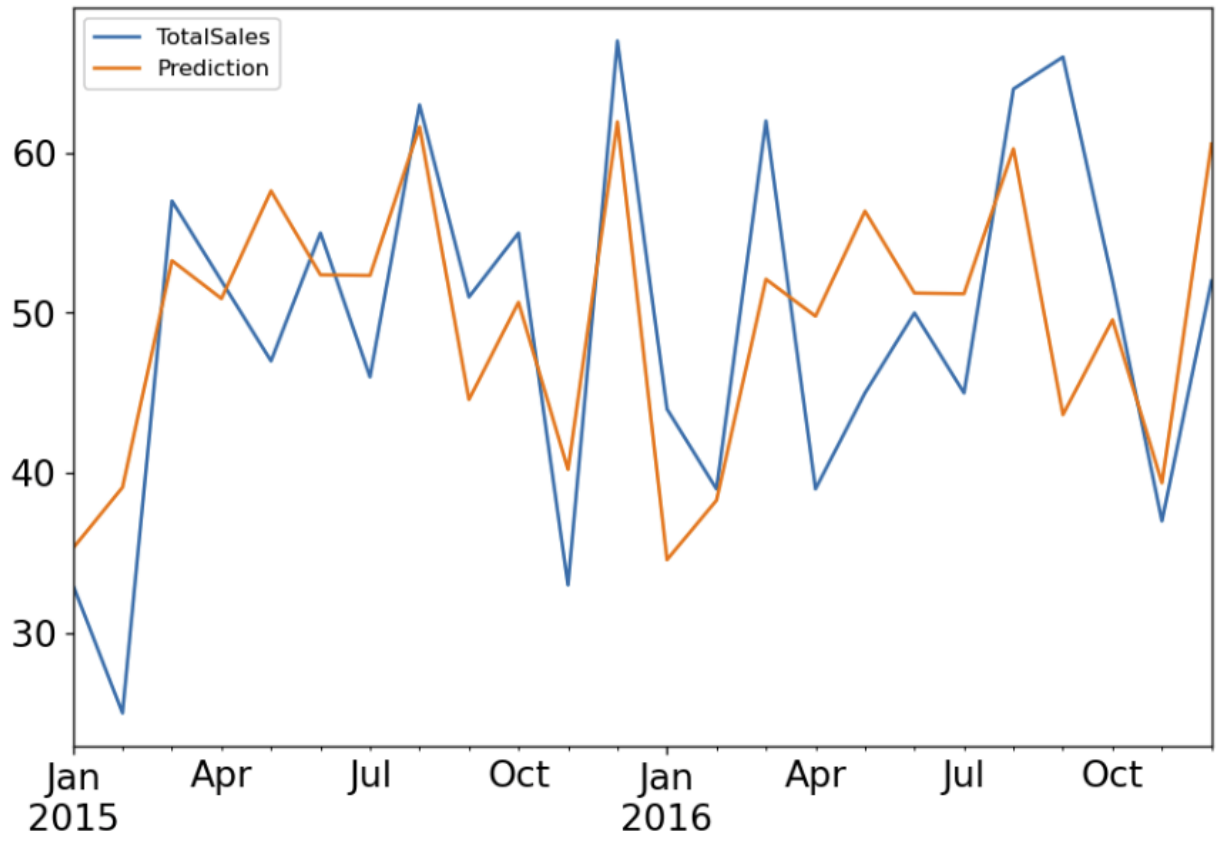
The best result out of the top three is an RMSE of about 8.69 degrees with the following configuration:

- **Trend:** Multiplicative
- **Damped:** True
- **Seasonal:** Multiplicative
- **Seasonal Periods:** 12
- **Box-Cox Transform:** True
- **Remove Bias:** False

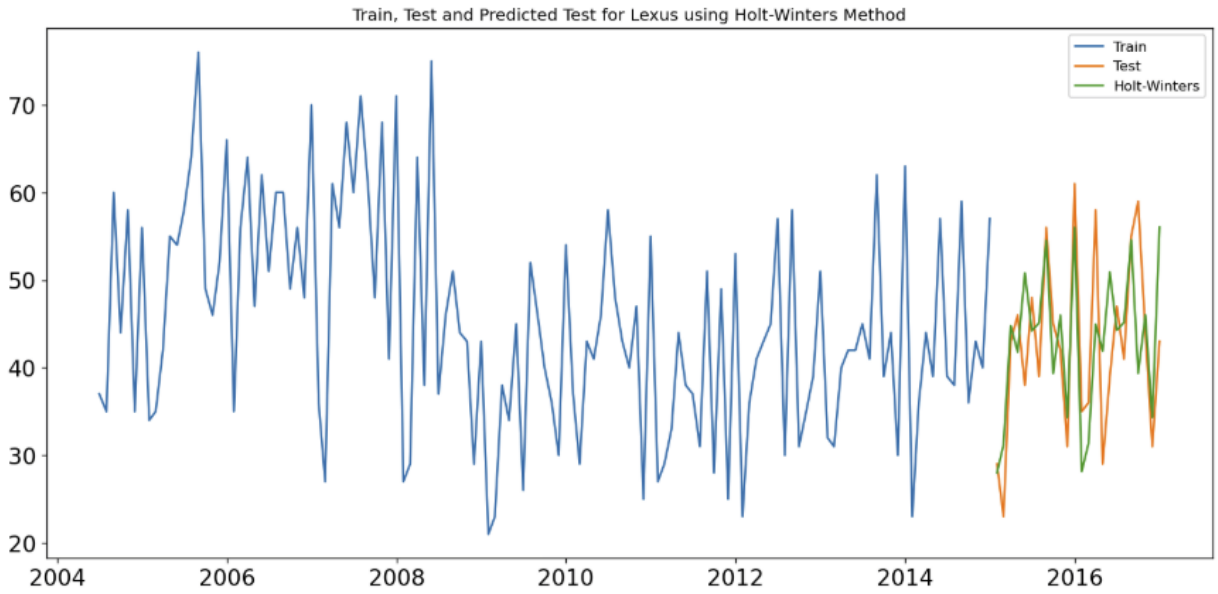
Validating Holt-Winter's Predictions

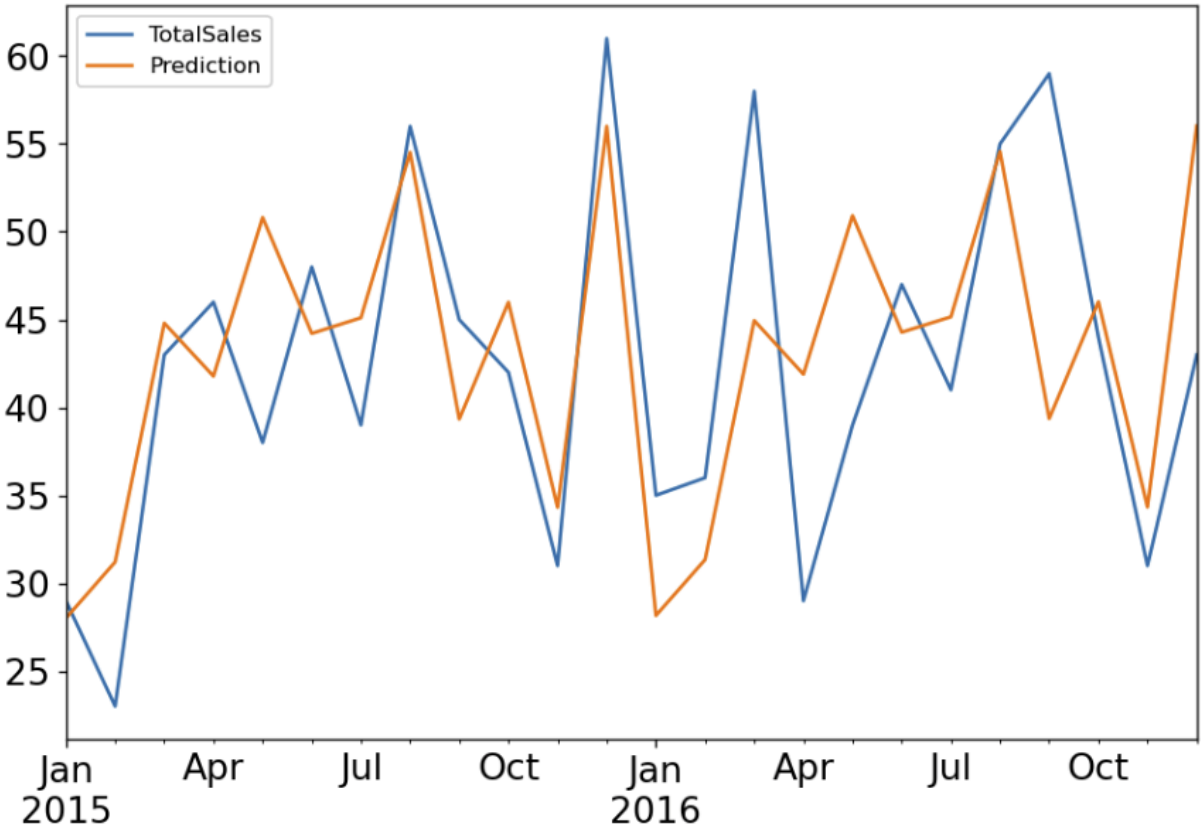
Dealership Prediction





Lexus Prediction





Model Evaluation

Dealership Evaluations

	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

Out of the four baseline methods explored on the total dealership sales dataset, we can see the Holt-Winter's method performed the best with a RMSE of 8.14 and a MAPE of around 14.2% which implies the model is about 85.8% accurate in predicting the test set observations.

Lexus Evaluations

	Method	RMSE	MAPE
0	Simple Average	10.380000	22.993712
0	Naive	17.710000	43.719574
0	SARIMAX	8.428568	17.403573
0	SARIMAX_autoc	8.161580	16.316625
0	Holt-Winters	7.993429	15.790187

Out of the four baseline methods explored on the exclusive Lexus sales dataset, we can see again the Holt-Winter's method performed the best with a RMSE of 7.99 and a MAPE of about 15.8% which implies the model is about 84.2% accurate in predicting the test set observations.

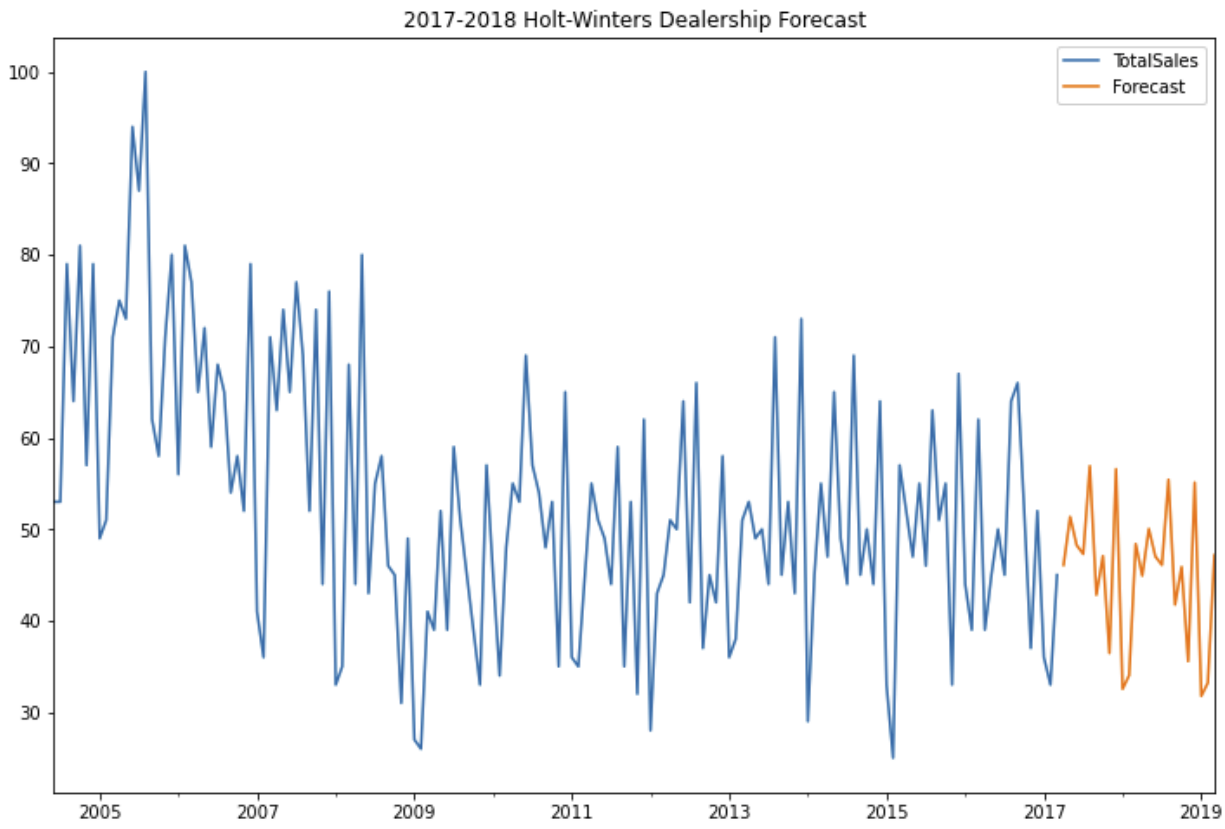
6.4 Preprocessing and Training Summary

As a result of our model evaluations, we now know that the Holt-Winter's model was the most accurate in predicting the test set observations. Now, we will save the Holt-Winter's model for both datasets as our final machine learning models. By forecasting these models in the next section we will be able to examine their predictions for future sales and answer the initial business problem proposed in our initial 'Data Wrangling' section.

7 Machine Learning Model Forecast

Final Model: Holt-Winter's Model

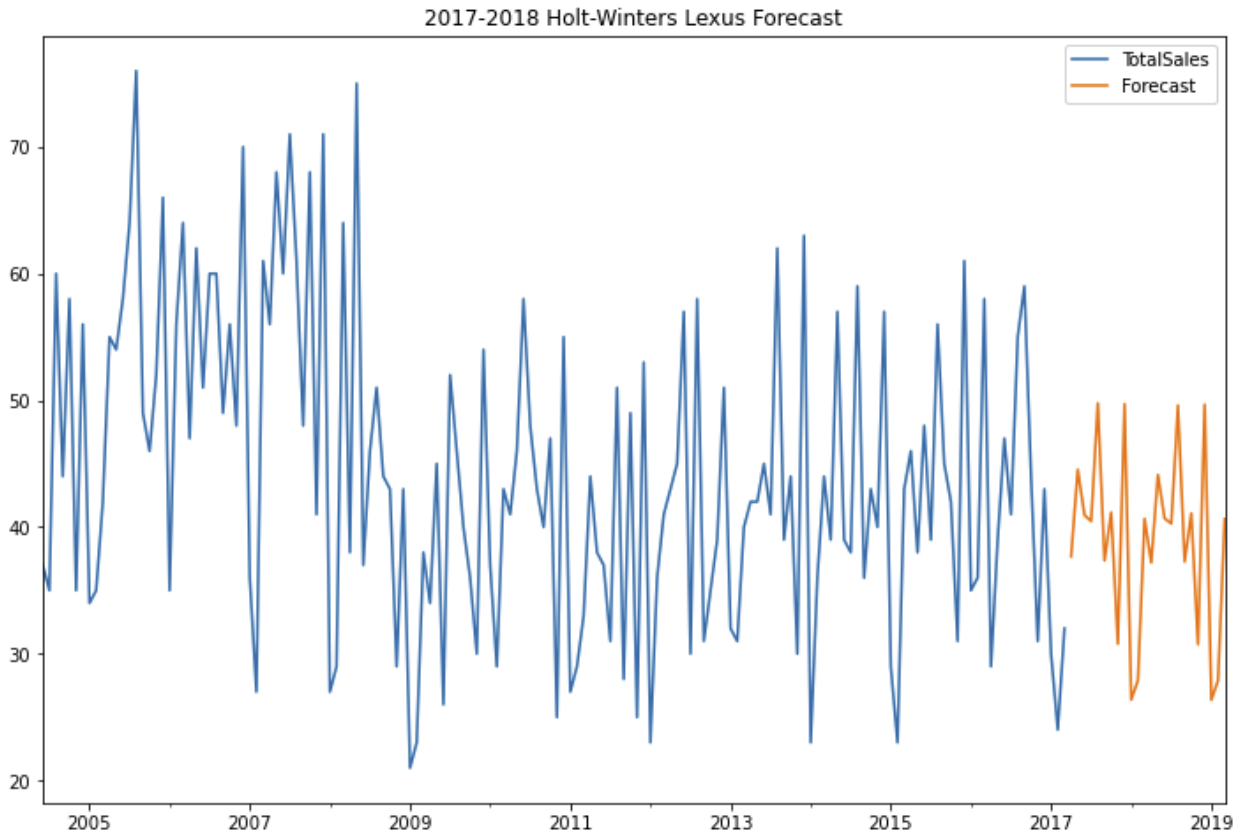
Dealership Forecast



Dealership forecasted sales for the next 24 months (after March 2017):

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

Lexus Forecast



Lexus forecasted sales for the next 24 months (after March 2017):

2017-04-30	37.685271
2017-05-31	44.538692
2017-06-30	40.947274
2017-07-31	40.462249
2017-08-31	49.770409
2017-09-30	37.357771
2017-10-31	41.156565
2017-11-30	30.779751
2017-12-31	49.704532
2018-01-31	26.376577
2018-02-28	27.905025
2018-03-31	40.655722
2018-04-30	37.197778
2018-05-31	44.122344
2018-06-30	40.679635
2018-07-31	40.275721
2018-08-31	49.604163
2018-09-30	37.272820
2018-10-31	41.089646
2018-11-30	30.745829
2018-12-31	49.663321
2019-01-31	26.362443
2019-02-28	27.894378
2019-03-31	40.644168

7. Suggested Improvement

There are several ways to improve performances in identifying ….

8. Project Summary

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