Unit sales Predictions

Capstone 1 Report

Immanuel Umenei

https://github.com/Mokonzi4n/Capstone_Project_1

Background: Solar Industry Fuse Circuit Protection requirement

- The solar industry has grown exponentially from 40GW in 2010 - 531GW in 2018 cumulatively. It is projected to reach 1.27 TW in 2022
- The energy produced by the sun is converted to electrical energy by solar panels. These are connected in series to attain the required voltage and in parallel to provide the required current.
- At the levels of parallel connections, fuses are used to protect the panels. Additionally, fuses are needed to protect wire that is used to carry current from the points of parallel connection to the inverter where this current is converted from direct current (DC) to alternating current (AC).
- This implies Utility scale solar installations would require thousands of fuses per installation. For instance a 1MW installation will require approximately 1300 fuses.

Problem

- The solar industry is project based, seasonal and policy driven. Slight changes in policy like increase in tariffs or changes in Federal Incentive Tax Credits completely change the annual expectations of the market. Making it very erratic and unpredictable.
- Additionally the project length and time to fuse requirement varies greatly from one project to another. And for most projects, by the time the fuses are needed in the field, the fuse manufacturers are required to respond within 2-4 weeks else would lose tens of thousands of dollars to competition having fuses on the shelve.
- The normal lead time for 5000 fuses is 5-6 weeks if there are no raw materials already purchased or finished goods on the shelf. This makes planning and manufacturing of fuses to meet customer needs highly dependent on customer input and forecasts.

Project Motivation: Forecasting Demand

- Demand forecasting is the estimation of the probable demand of a product or service in the future and it is based on the analysis of passed demand and aspects of the market for the specific product or service.
- Planners and Production Managers of Manufacturing companies are faced with the difficulty year in year out to plan and budget for raw materials to be used to build finished goods for the following year.
- They have to estimate or plan according to forecasts provided by the sales and product management teams, historical and market data.
- Inaccurate estimates or information from the sales teams lead to under stocking or overstocking. Hence increased operating costs
- This is useful in making business decisions related to sales, production, staff requirement etc. in especially new and emergent markets which are usually competitive and hard to predict.
- The effectiveness of business decisions is highly dependent on the accuracy of the information gathered with respect to the market and business. This accuracy is dependent on that of the demand forecasts especially when related to sales, production etc.
- Having accurate forecasts in the solar industry for fuse manufacturers will be very useful in helping with production planning, process selection, capacity planning, inventory management, facility layout planning. It will provide reasonable data for the organization's capital investment and expansion decision.
- Using Data Science techniques to implement models that analyze sales data and provide accurate or close to accurate predictions of monthly sales will be an invaluable tool in the hands of different functional teams within any business organization
- The main clients would be sales, production and planning teams of manufacturing industries that want to be leaner in their budgeting and planning for production, to minimize risks and losses due to under or over estimation of annual product sales.

Data Science Approach

- Export data from the SAP business operations software in an excel format.
- Wrangle the data in pandas to eliminate any bad and irrelevant data or outliers.
- Do an EDA on the data to determine if there are any trends by year, month, by region, and by SKU.
- Use 4 years (2014 - 2017) of data. Splitting it into a training and test data set in the ARIMA model
- Validate the ARIMA model on the 5th year (2018).
- Refine the model as needed until predicted sales data by the model is as close as possible to the actual 2018 sales

Data Set

- Fives years (2014 - 2018) of global sales units data by month, by quarter and by region of all product SKUs sold into the renewable energy industry specifically PV solar.

Data Wrangling

- Tidy Data Format:
  Before pulling the data into a .csv file, features of the business operating software were used to arrange the data in a tidy format where columns represent separate variables and each row represents individual observations.

	Quote	End Customer	Product Hierarchy	Product Family	Geographic Region	Sold-to party	Distributor	Part Number	Month of Year	Quarter of Year	Day	Net Inv Qty	Net Inv-$
0	D05013	*SANMINA/ARROW YRLY 2013	PPF650NA09 ZZ	Midgit KLKD	US/CAN	306110	ARROW PEMCO GROUP	KLKD030.T	Period 01 2014	20141	2014-01-01 00:00:00	NaN	-25086.51
1	JPIS012	SUN-WA TECHNOS(DAIHEN)	PPF650NA09 ZZ	Midgit KLKD	ASIA	405253	FUJIX CO., LTD.	KLKD005.T	Period 01 2014	20141	2014-01-01 00:00:00	200.0	558.93

- Inconsistent Column Names: Converted "Month of Year" rows into date time format and renamed it "Month" and "Quarter of Year" into "Quarter" with alphanumeric representations (i.e. Q1, Q2, Q3, Q4).

	End Customer	Product Hierarchy	Product Family	Geographic Region	Sold-to party	Distributor	Part Number	Month	Year	Quarter	Day	Net Inv Qty	Net Inv-$
0	*SANMINA/ARROW YRLY 2013	PPF650NA09 ZZ	Midgit KLKD	US/CAN	306110	ARROW PEMCO GROUP	KLKD030.T	1	2014	Q1	2014-01-01 00:00:00	NaN	-25086.51
1	SUN-WA TECHNOS(DAIHEN)	PPF650NA09 ZZ	Midgit KLKD	ASIA	405253	FUJIX CO., LTD.	KLKD005.T	1	2014	Q1	2014-01-01 00:00:00	200.0	558.93

Data Wrangling

- Missing data:

Ran the .info() attribute and realized the "Net Inv Qty column had some missing data. There are 3764 missing values on the Qty column. We apply the dropna() method to drop all missing value rows

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 29916 entries, 0 to 29915
Data columns (total 13 columns):
End Customer          29916 non-null object
Product Hierarchy     29916 non-null object
Product Family        29916 non-null object
Geographic Region     29916 non-null object
Sold-to party         29916 non-null int64
Distributor           29916 non-null object
Part Number           29916 non-null object
Month                 29916 non-null int64
Year                  29916 non-null int64
Quarter               29916 non-null object
Day                   29916 non-null object
Net Inv Qty           26152 non-null float64
Net Inv-$             29916 non-null float64
dtypes: float64(2), int64(3), object(8)
memory usage: 3.0+ MB
```

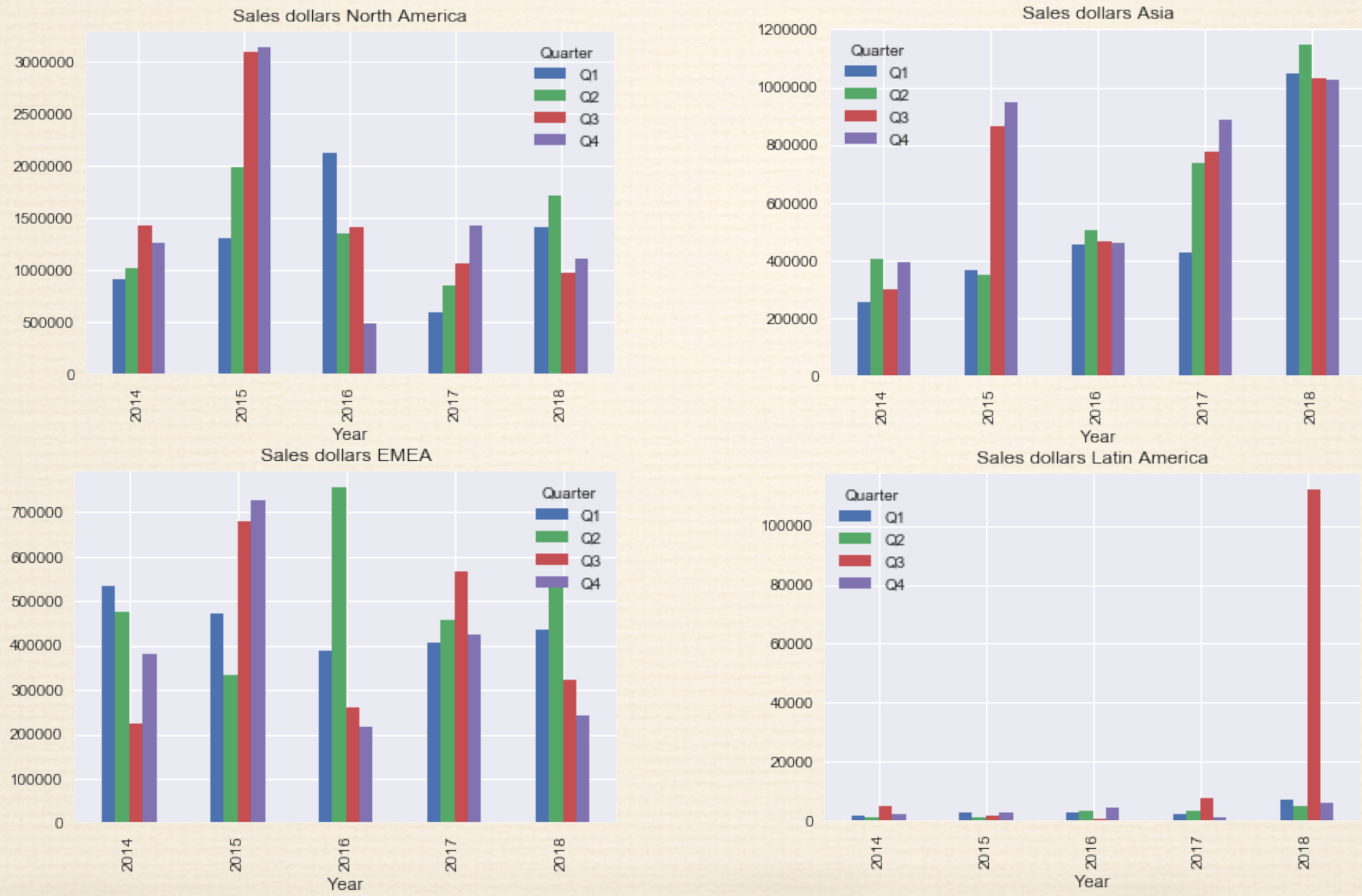
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 26152 entries, 1 to 29915
Data columns (total 13 columns):
End Customer          26152 non-null object
Product Hierarchy     26152 non-null object
Product Family        26152 non-null object
Geographic Region     26152 non-null object
Sold-to party         26152 non-null int64
Distributor           26152 non-null object
Part Number           26152 non-null object
Month                 26152 non-null int64
Year                  26152 non-null int64
Quarter               26152 non-null object
Day                   26152 non-null object
Net Inv Qty           26152 non-null float64
Net Inv-$             26152 non-null float64
dtypes: float64(2), int64(3), object(8)
memory usage: 2.8+ MB
```

```
 10.0     5342
NaN       3764
 20.0     2377
 100.0    2101
 30.0     1163
Name: Net Inv Qty, dtype: int64
```

```
10.0     5342
20.0     2377
100.0    2101
30.0     1163
50.0     1134
Name: Net Inv Qty, dtype: int64
```

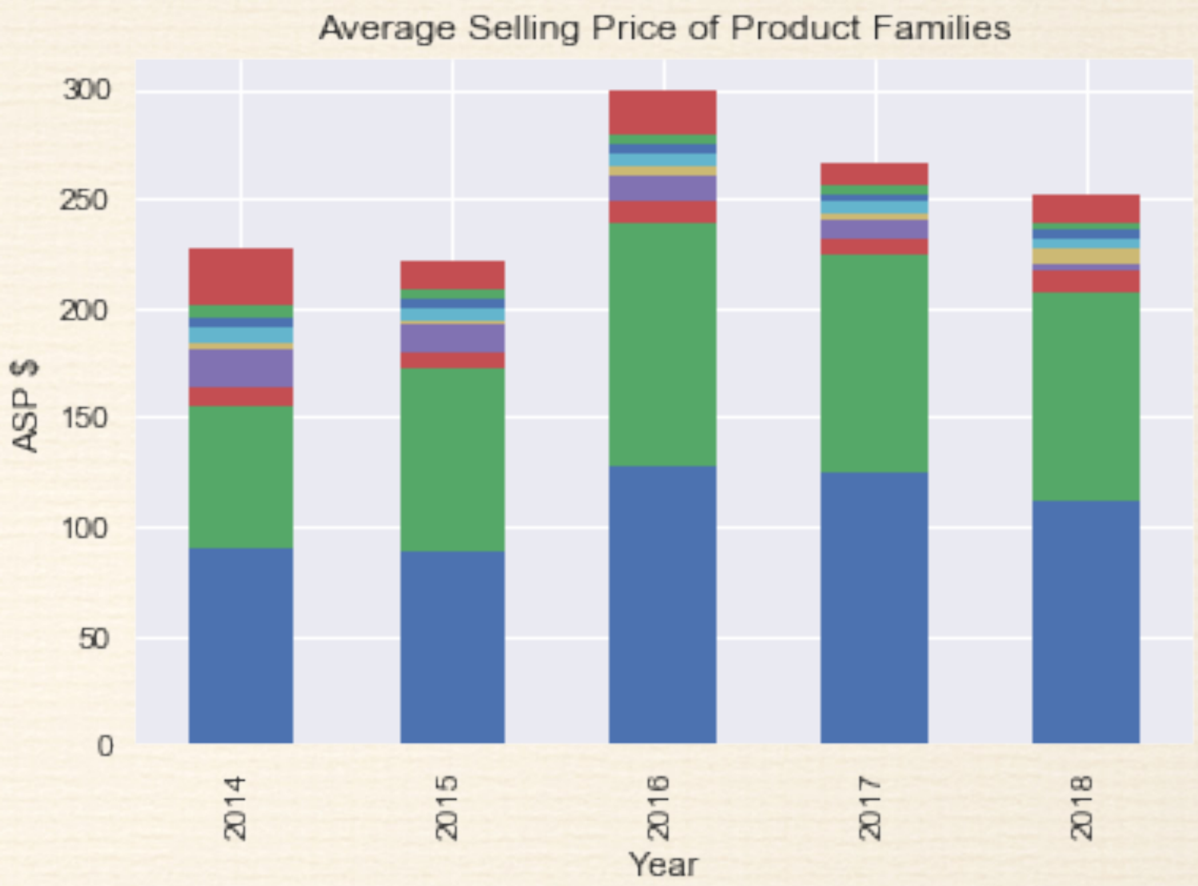
Data Story Telling

- Regional Comparisons by Quarter:
  - The strongest markets are the US and Asia followed by Europe and then Latin America
  - 2015 was a great year for both US and Asia markets but 2016 saw a steep decline in these markets
  - Since 2014 Europe has been on a consistent decline in units sold while Latin America has been growing slowly having a remarkable quarter in Q3 of 2018



Data Story Telling

- Product Family Comparison:
  - The average selling prices overall have fluctuated but have seen a drop over the years.
  - Comparing only product families sold across the 5 years we have that 2016 was the highest year of ASPs but has seen a decline across all product families since then

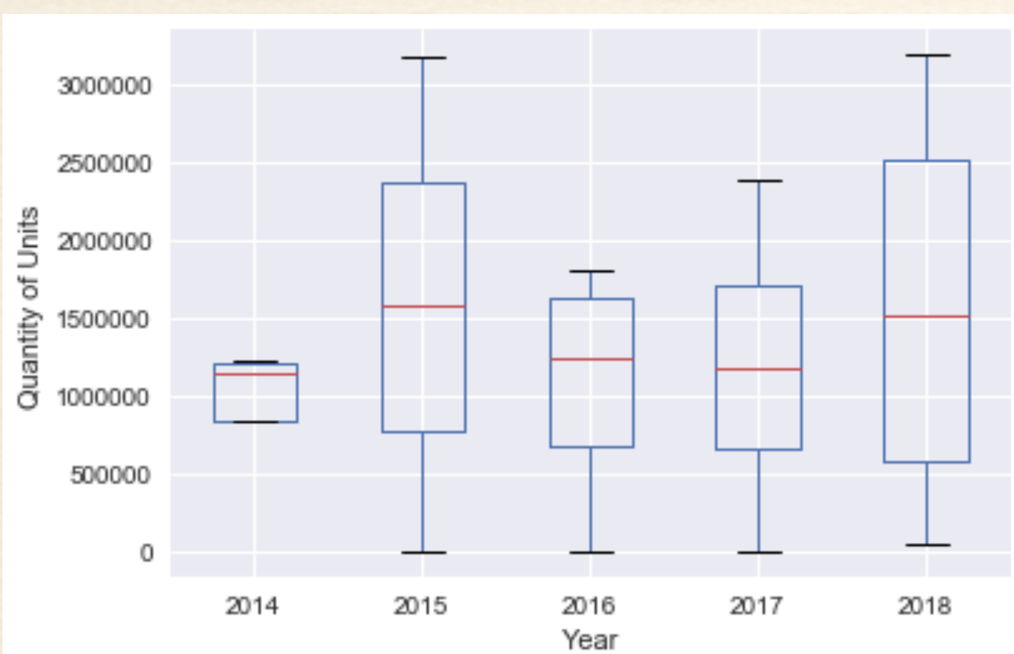


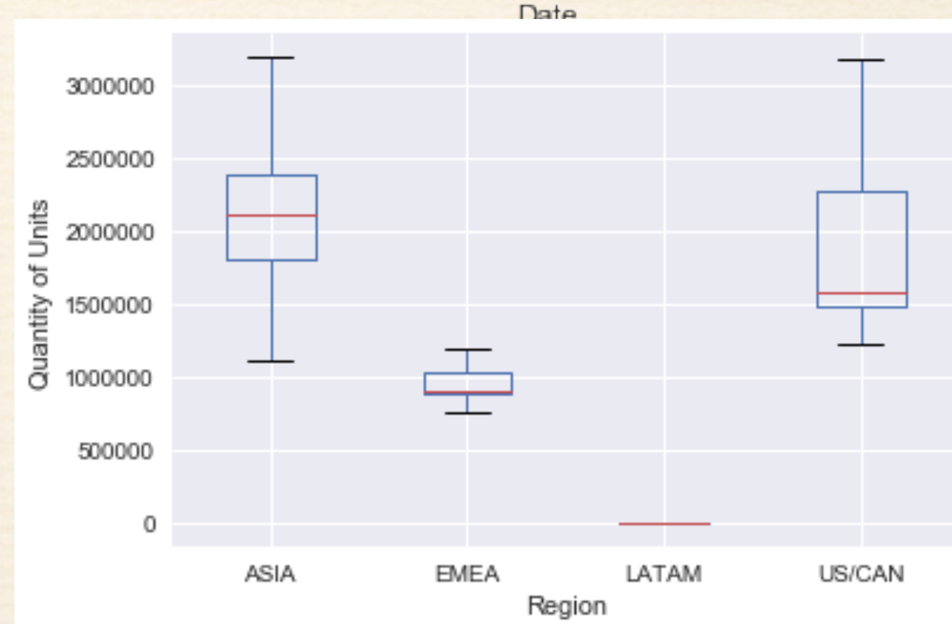
Exploratory Data Analysis

- The main objective is to analyze the data to understand any trends and patterns and to be able to use the first four years to forecast the fifth year.

- Looking at the figure it is clear that the best years so far have been 2015 and 2018. The most variability in terms of quantity sold across the 4 regions is in 2018. The biggest markets have been Asia and the US

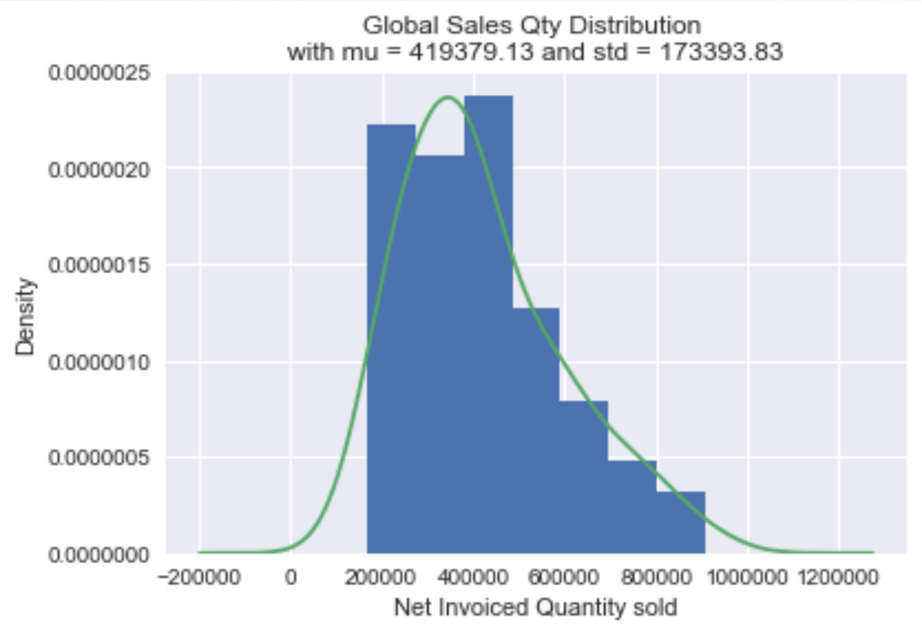
	Net Inv Qty				
Year	**2014**	**2015**	**2016**	**2017**	**2018**
count	4.000000e+00	4.000000e+00	4.000000e+00	4.000000e+00	4.000000e+00
mean	8.824770e+05	1.582397e+06	1.070201e+06	1.188924e+06	1.566689e+06
std	5.887321e+05	1.369326e+06	8.072041e+05	1.002952e+06	1.426316e+06
min	2.512000e+03	2.460000e+03	3.296000e+03	6.870000e+03	4.747500e+04
25%	8.312208e+05	7.788308e+05	6.765260e+05	6.623700e+05	5.764935e+05
50%	1.152010e+06	1.573098e+06	1.240136e+06	1.179558e+06	1.516326e+06
75%	1.203266e+06	2.376664e+06	1.633811e+06	1.706111e+06	2.506521e+06
max	1.223376e+06	3.180932e+06	1.797236e+06	2.389709e+06	3.186628e+06

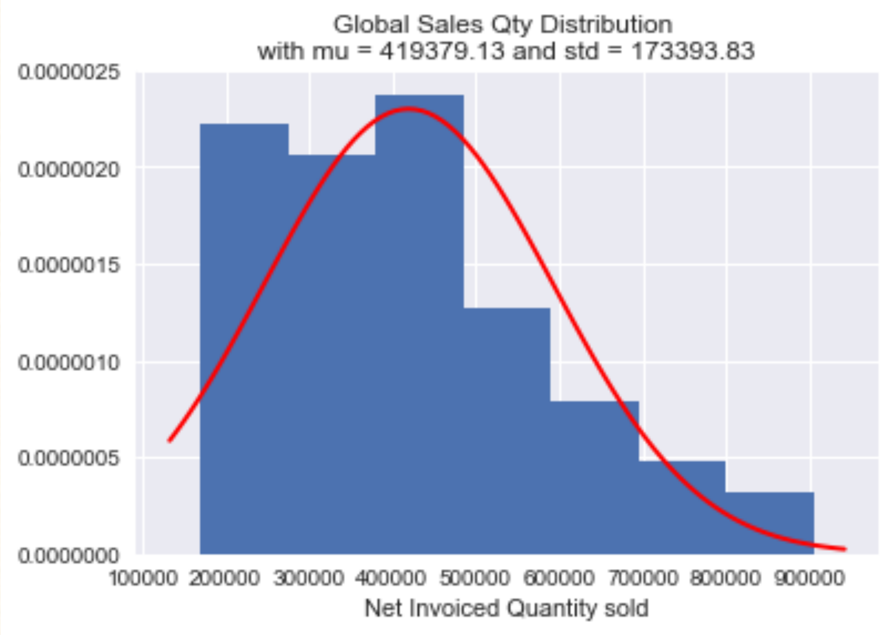


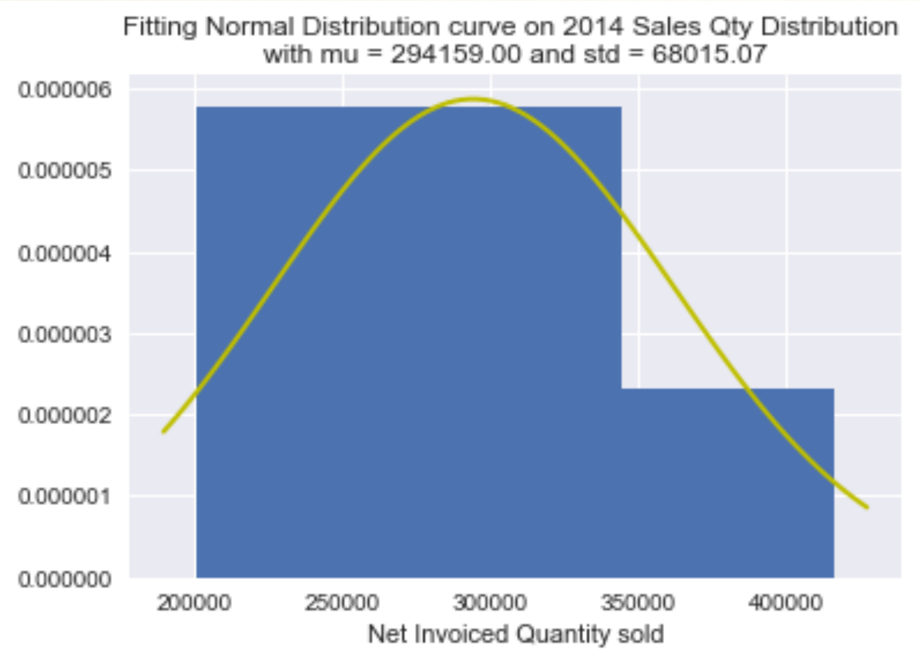


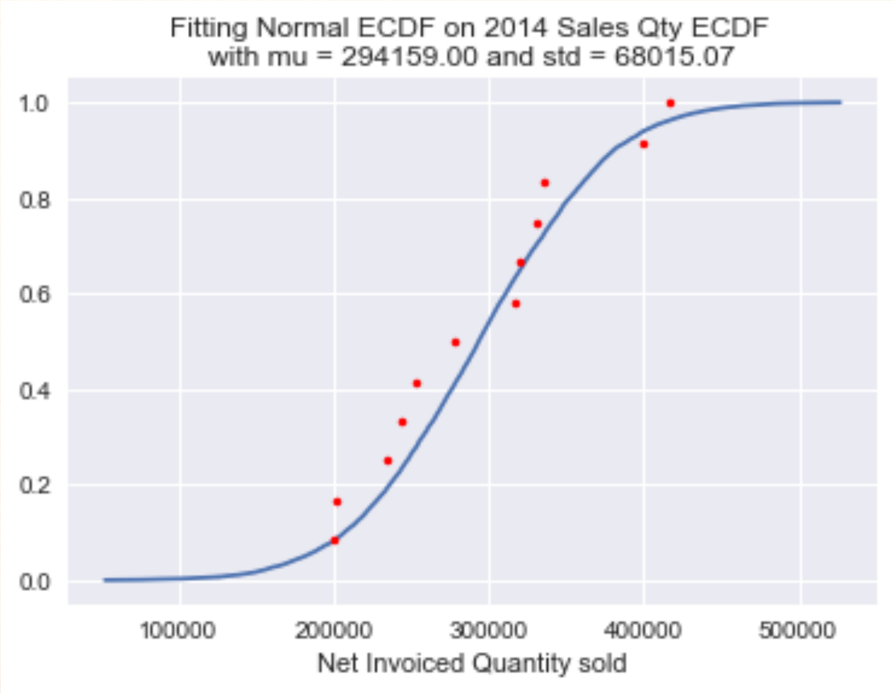
Exploratory Data Analysis

- Global Quantity Sales Distribution: Doing a normality test and plotting the distribution of the total quantity sold globally for the 5 years of data, it gives a right skewed distribution. Which indicates that there are fewer months in which the very high quantities were bought.



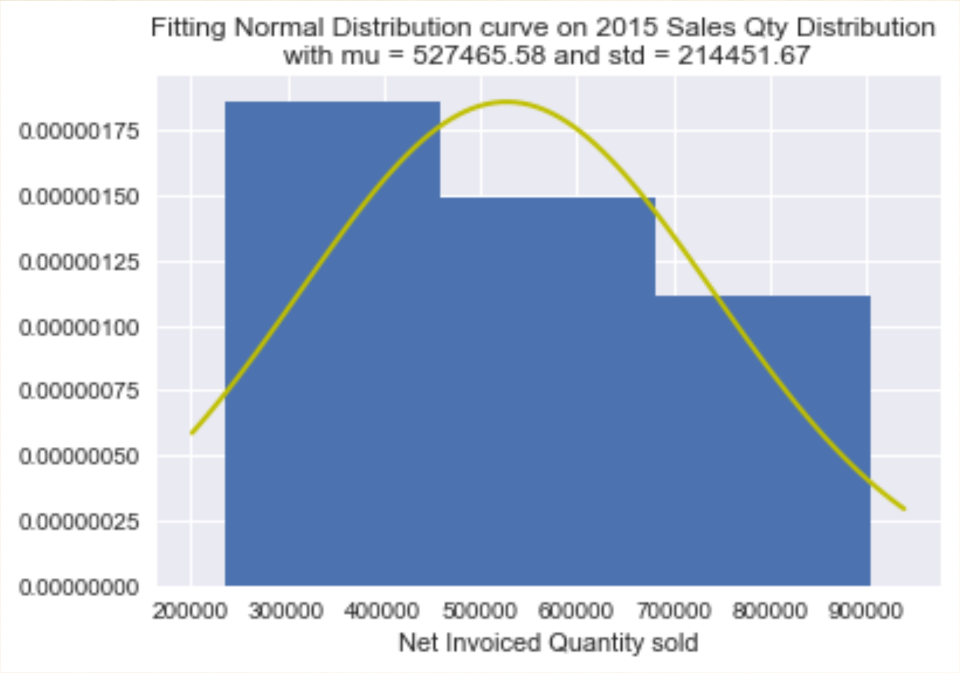


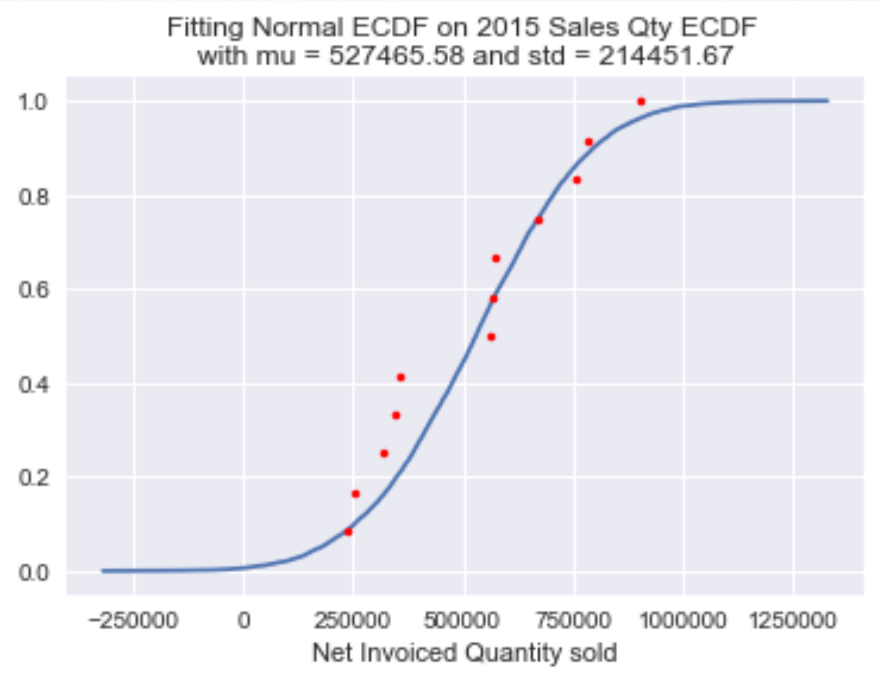


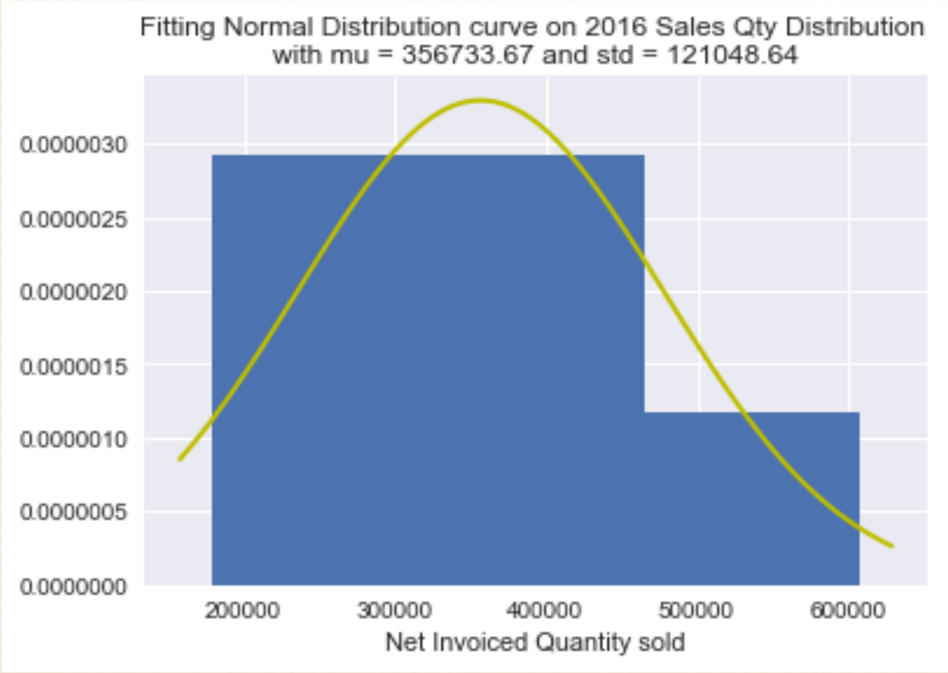


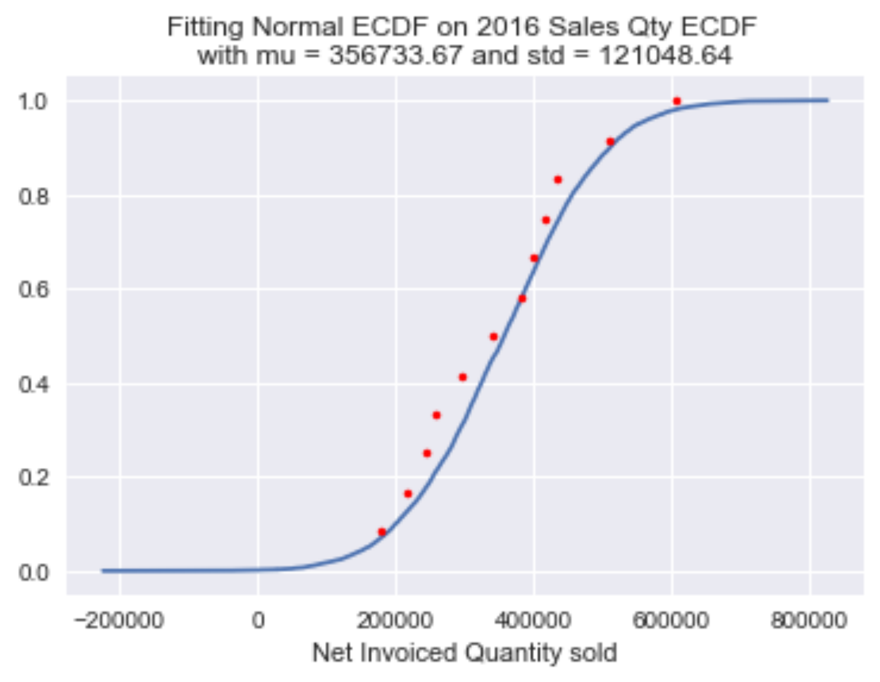
Exploratory Data Analysis

- The data varies greatly from year to year and there does not seem to be a specific trend in the sales quantities. The distribution of the sales quantity by year also is not normal..

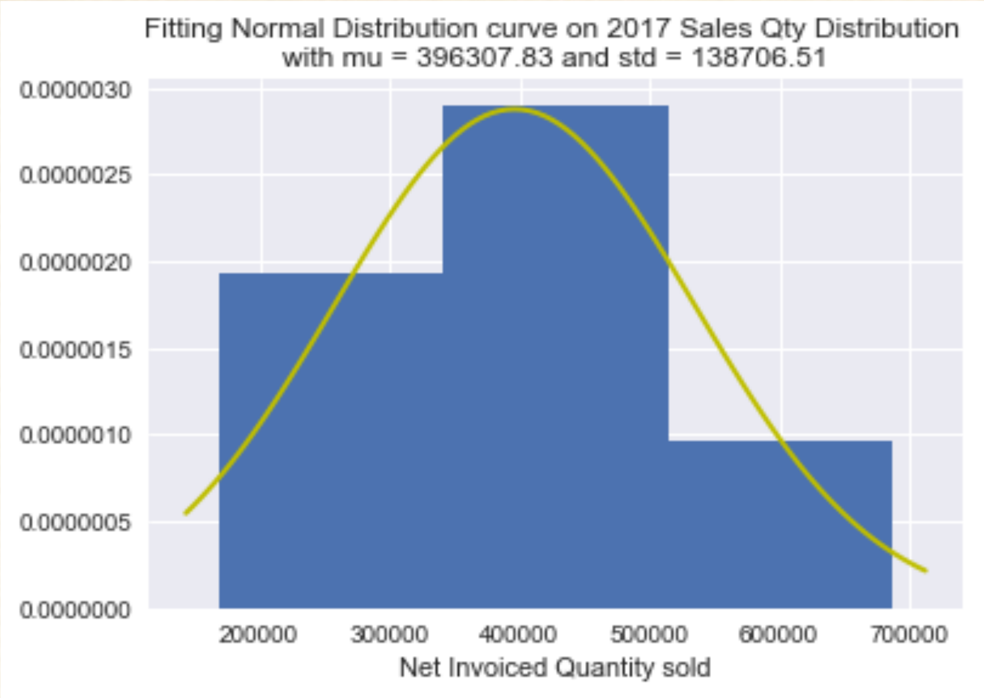


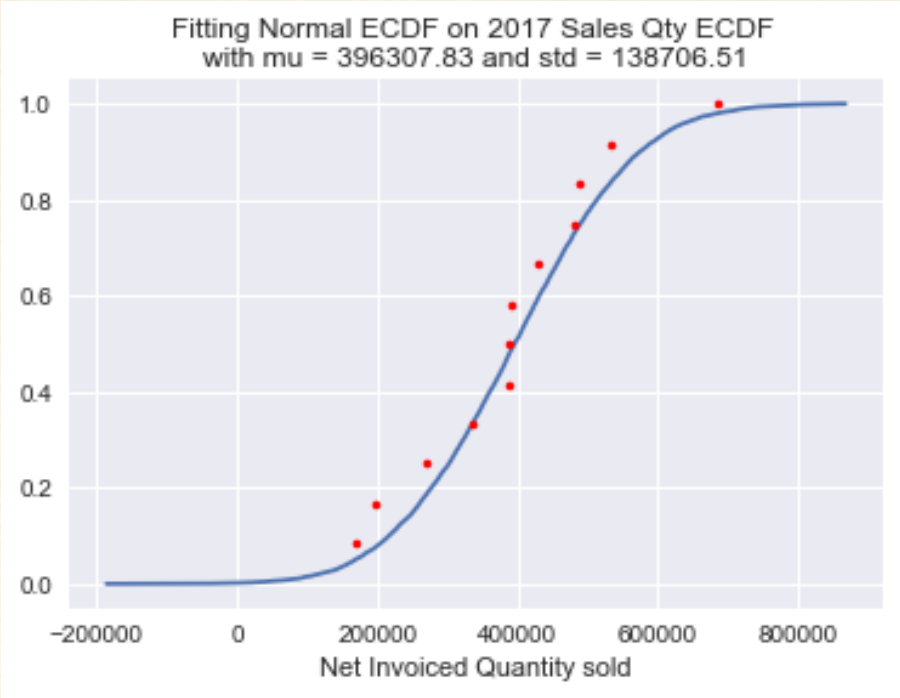


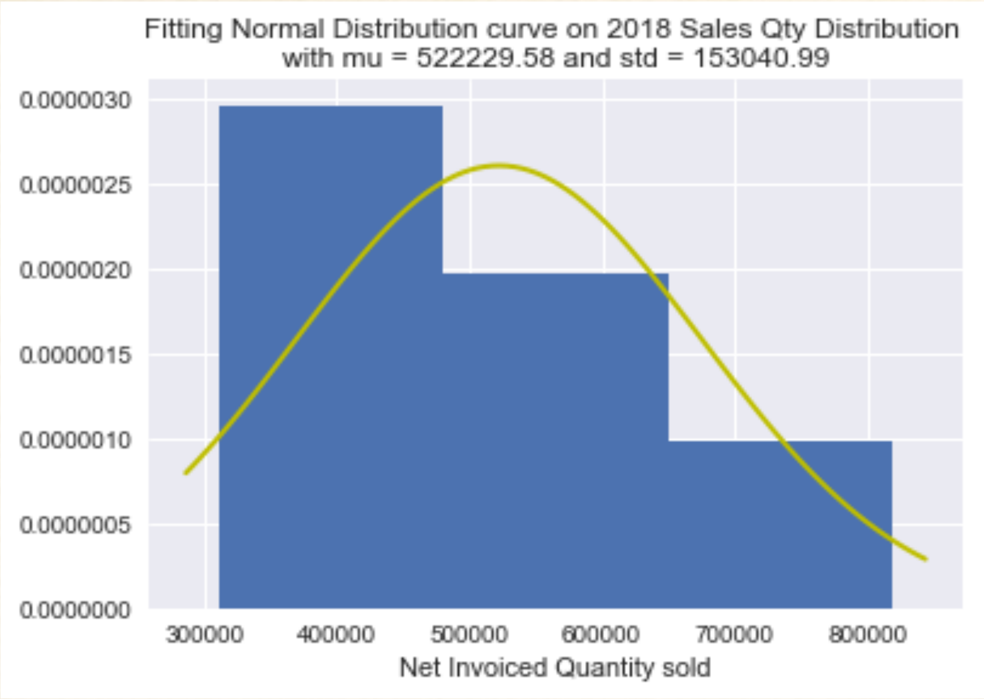


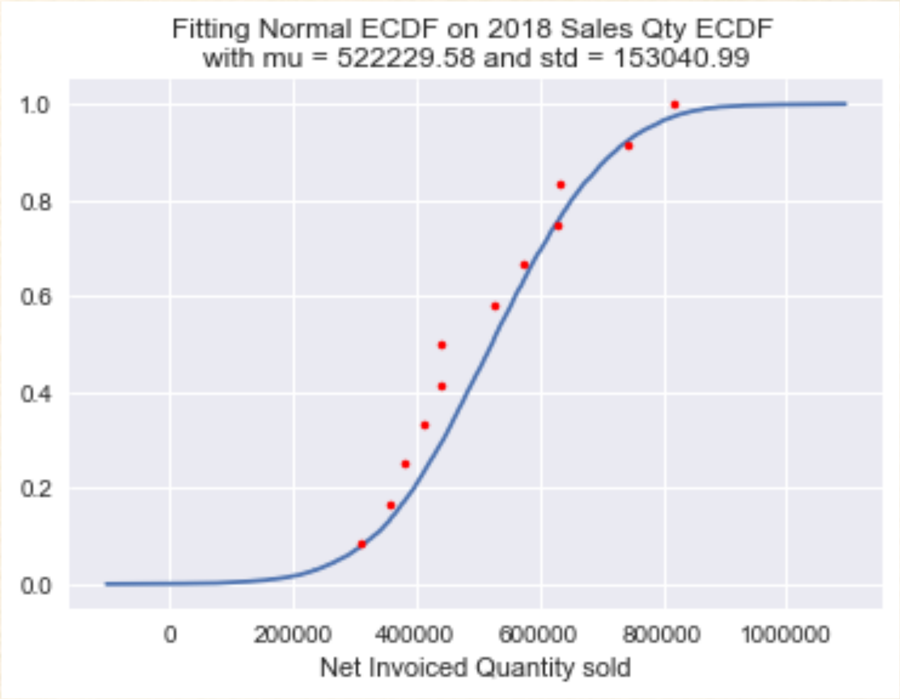


Exploratory Data Analysis









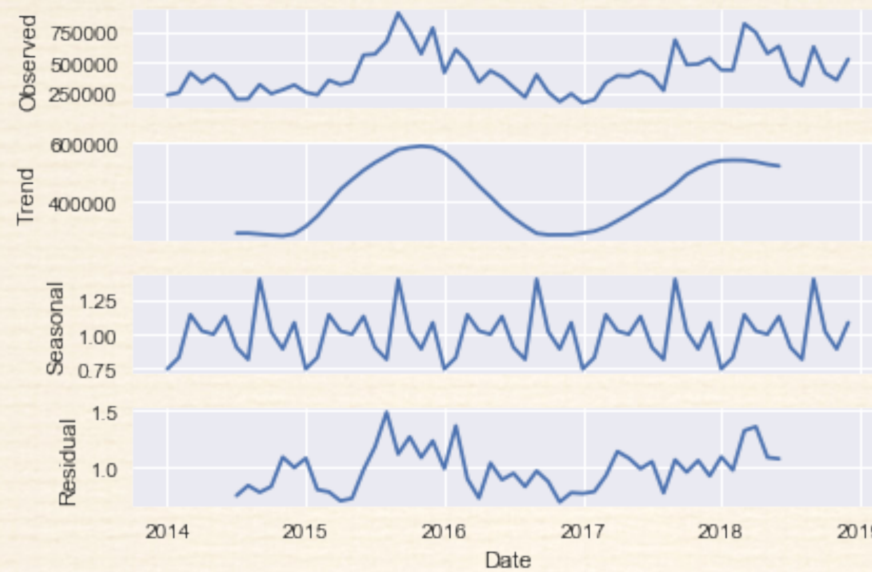
- To better forecast the sales quantities, we will need to account for the variability and seasonality of the data. One of the ways to do this is to use an ARIMA model.

Machine Learning Processes

- The ARIMA model is used
- The data set is split from 2014 - 2017 as train and test data sets and 2018 is held out as validation set.
- Use the train and test data sets to train and fine tune the model to get optimal hyper parameters of the model. These are chosen based on their Root Mean Square Error values.
- Validate the model by fitting it using the determined hyper parameters on the validation data sets to predict the outcomes of the this data set.

Data Analysis

- As shown in the EDA above and the seasonal decomposition below the series is not stationary



- Stationarize series using differencing and use Augmented Dickey-Fuller (ADF) test to confirm that the series is stationary.

- The results show that the test statistic value -3.208282 is smaller than the critical value at 5% of -2.968. This suggests that we can reject the null hypothesis with a significance level of less than 5%

```
ADF Statistic: -3.208282
p-value: 0.019510
Critical Values:
    1%: -3.679
    5%: -2.968
    10%: -2.623
```

Grid Search Hyperparameter (p,d,q) Tuning

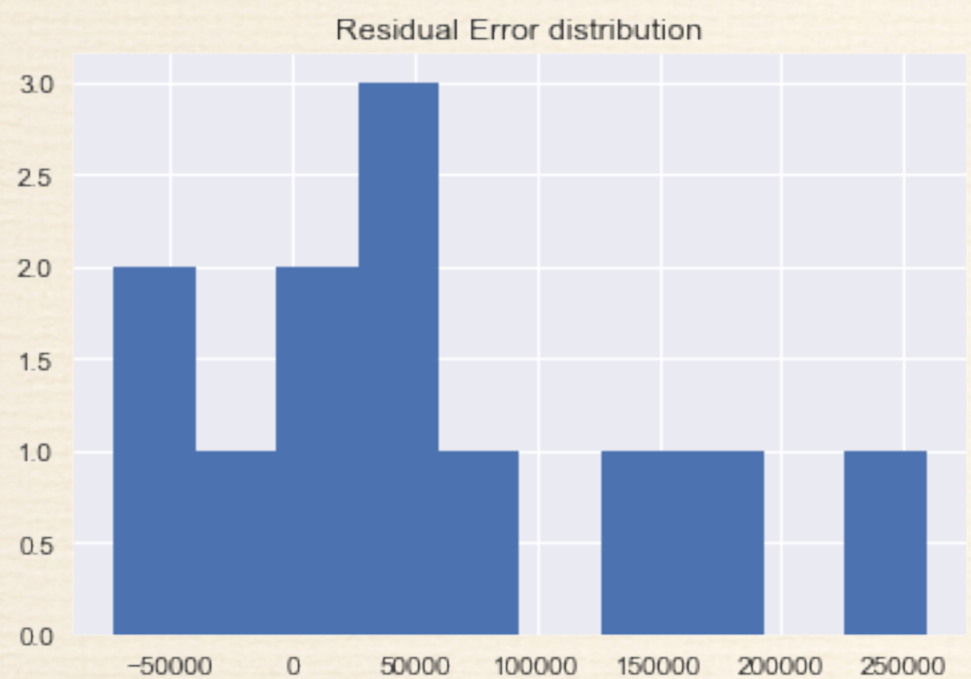
- Number of AR (AutoRegressive) terms parameter p (0-6)
- Number of non-seasonal difference terms parameter d (0-2)
- Number of MA (Moving Average) terms (0-6)
- 147 possible combinations are iterated over and fitted on the model.
- Best (p,d,q) order is (3,0,2) based on the RMSE = 109681.34

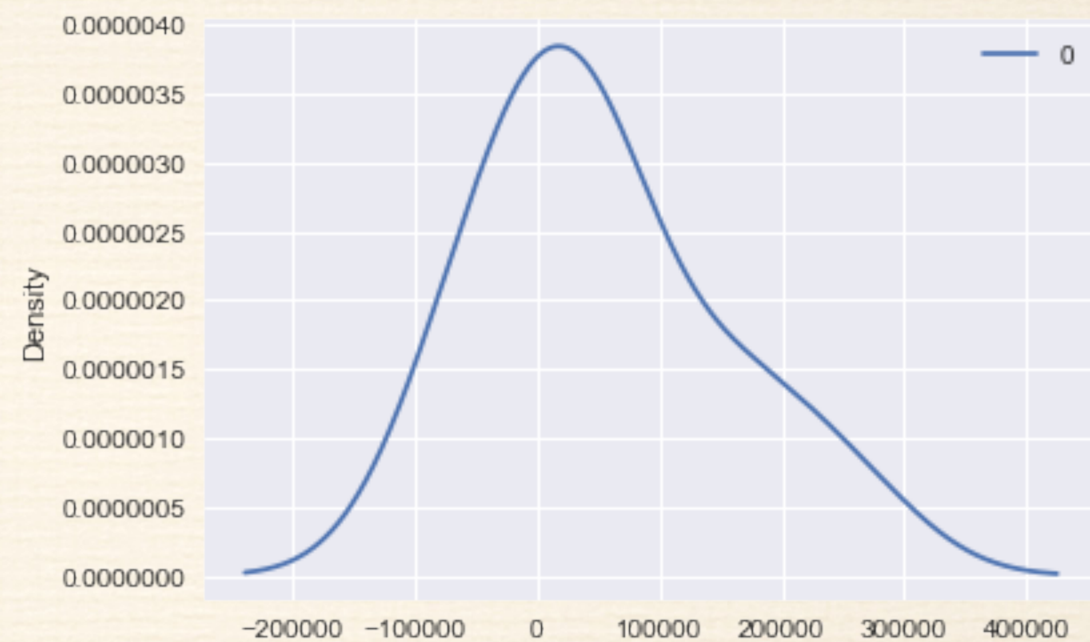
```
ARIMA(0, 0, 1) RMSE=176715.312
ARIMA(0, 0, 2) RMSE=154134.801
ARIMA(0, 1, 1) RMSE=153774.357
ARIMA(0, 1, 2) RMSE=131968.388
ARIMA(0, 1, 3) RMSE=151724.918
ARIMA(0, 2, 1) RMSE=163022.663
ARIMA(1, 0, 0) RMSE=147474.555
ARIMA(1, 0, 1) RMSE=144783.848
ARIMA(1, 0, 2) RMSE=115516.592
ARIMA(1, 0, 3) RMSE=151316.670
ARIMA(1, 1, 0) RMSE=150941.263
ARIMA(1, 2, 0) RMSE=175838.561
ARIMA(1, 2, 1) RMSE=149155.371
ARIMA(1, 2, 2) RMSE=151142.707
ARIMA(2, 0, 0) RMSE=142956.671
ARIMA(2, 0, 1) RMSE=142924.275
ARIMA(2, 0, 2) RMSE=138858.046
ARIMA(2, 1, 0) RMSE=149306.167
ARIMA(2, 1, 1) RMSE=142318.294
ARIMA(2, 2, 0) RMSE=135387.738
ARIMA(2, 2, 1) RMSE=137041.983
ARIMA(3, 0, 0) RMSE=140175.384
ARIMA(3, 0, 2) RMSE=109681.338
ARIMA(3, 1, 0) RMSE=130418.947
ARIMA(3, 2, 0) RMSE=137650.629
ARIMA(3, 2, 1) RMSE=138385.060
ARIMA(3, 2, 2) RMSE=144995.899
ARIMA(4, 0, 0) RMSE=111325.047
ARIMA(4, 0, 1) RMSE=112626.271
ARIMA(4, 0, 2) RMSE=135771.483
ARIMA(4, 1, 0) RMSE=130890.913
ARIMA(4, 1, 1) RMSE=125611.932
ARIMA(4, 2, 0) RMSE=142953.600
ARIMA(4, 2, 1) RMSE=143084.655
ARIMA(5, 0, 1) RMSE=113795.698
ARIMA(5, 1, 0) RMSE=130400.244
ARIMA(5, 1, 1) RMSE=131604.074
ARIMA(5, 1, 2) RMSE=159613.165
ARIMA(5, 2, 0) RMSE=139256.652
ARIMA(5, 2, 1) RMSE=137821.443
ARIMA(6, 1, 0) RMSE=131934.940
ARIMA(6, 1, 1) RMSE=128567.032
ARIMA(6, 2, 0) RMSE=137337.504
ARIMA(6, 2, 1) RMSE=144402.803
Best ARIMA(3, 0, 2) RMSE=109681.338
```

Residual Error Review

- The residual error distribution of the data set is approximately normal but not centered around zero as shown by the plot and summary statistics.

- Ideally the distribution of residual errors of a model should be normal around a mean of zero to indicate a good fit for the model



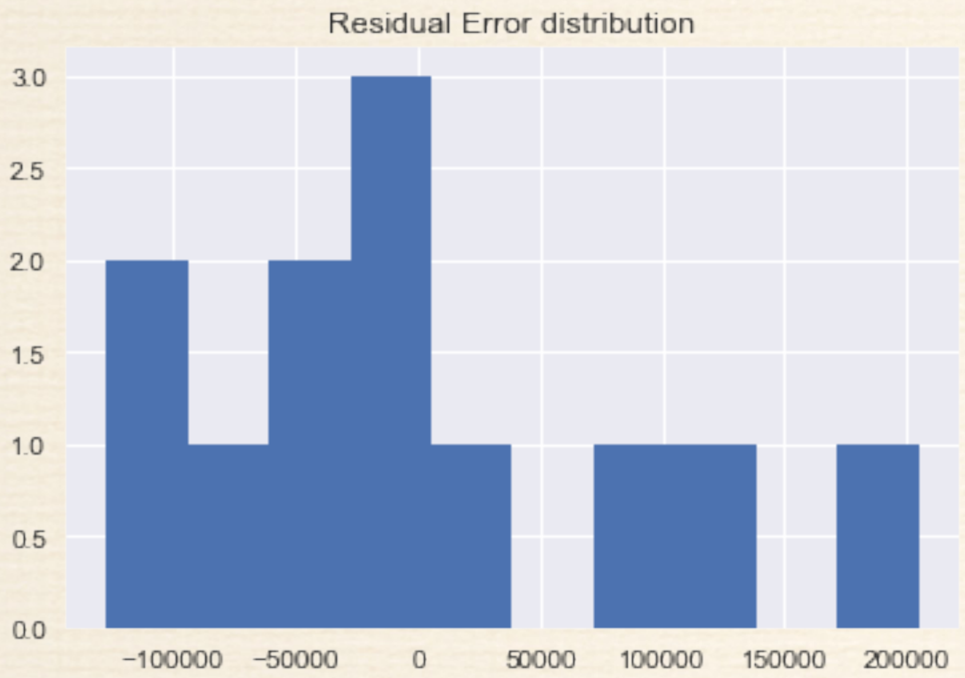


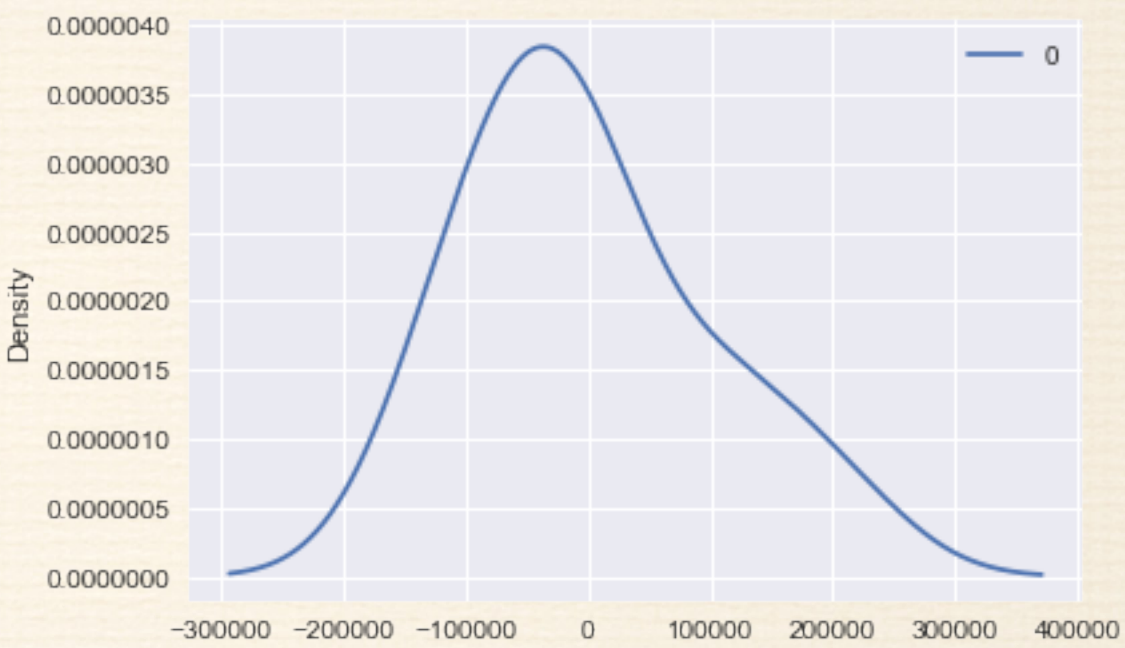
count	12.000000
mean	54396.846787
std	99476.697649
min	-72876.932260
25%	-4378.227423
50%	41685.485194
75%	95578.695404
max	259113.460964

Residual Error Review

- To center the distribution around zero we bias-correct the prediction with the residual error mean of 54396.85 added to each prediction

- Making this adjustment gives an improved distribution centered very close to zero (shown below) with an improved RMSE: 95241.687 from RMSE : 109681.3.





count	1.200000e+01
mean	4.008907e-07
std	9.947670e+04
min	-1.272738e+05
25%	-5.877507e+04
50%	-1.271136e+04
75%	4.118185e+04
max	2.047166e+05

Model Validation

- The trained model with optimal hyper-parameters is used in a rolling-forecast manner updating the model for each time step. This is used to predict 12 month of sales
- Predictions vs Observed are as shown below:

	Expected Values	Predicted Values	Forecast Error
2018-01-01	439508.0	445130.90	5622.90
2018-02-01	438470.0	502603.05	64133.05
2018-03-01	818048.0	629047.66	-189000.34
2018-04-01	744324.0	795527.38	51203.38
2018-05-01	573164.0	920468.51	347304.51
2018-06-01	634356.0	661309.71	26953.71
2018-07-01	379933.0	528069.18	148136.18
2018-08-01	311773.0	310283.52	-1489.48
2018-09-01	628897.0	642599.62	13702.62
2018-10-01	414029.0	406589.58	-7439.42
2018-11-01	356277.0	372906.52	16629.52
2018-12-01	527976.0	393080.82	-134895.18

Model Evaluation

- Performance Measures:
  - Below are the performance measures calculated for this model predictions.

	Performance Measures	Value
0	Forecast Bias	28405.118
1	MAE	83875.856
2	MAPE	15.366
3	RMSE	130542.906

- The forecast bias is positive indicating the model's tendency to over forecast. This explains some of the high values in May and July in particular.
- The MAPE is 15.4% which indicates the model has about an 85% accuracy on quantity sold.
- From a sales, raw material and labor standpoint you would want to over-forecast by 15% than under-forecast.
- Overall it is pretty good model owing to the fact that we considered only historical data.

Recommendations

- The model should be used with the guidance that the predictions would have a tendency to over forecast.
- Analysis was done on the total number of units of all product types sold into the renewable industry. Analysis should be done on each product family independently for better outcome as other components potentially add noise to the analysis.
- Additional features that influence the market like amount of installed capacity and policy change indices could be introduced into the dataset to fine tune the predictions with machine learning techniques like Random Forest and Neural Networks.