Capstone Project 1: Milestone Report

Problem:
Planners and Product 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. There is a tendency to estimate or plan according to forecasts provided by the sales teams or based on the volume done in the previous year. This leads to situations where demand cannot be met in a timely manner due to understocking or where demand drops off and there is overstock on raw materials for a SKU that ends up not being demanded as forecasted. The latter leads to scrapping of finished goods and hence increased operating costs. The objective here is to analyze sales data with data science techniques, to predict with a greater amount of precision the number of SKUs that would be required in the following year. This provides the manufacturing teams with the SKU volume information necessary for more accurate raw materials purchases and proper production lines scheduling.

Client:

Manufacturing industries that want to be leaner in their budgeting and planning for production, to minimize losses due to under or over estimating specific products

Data Set:

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

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 of data to predict the monthly SKU requirements for the 5th year.

Data Wrangling

Cleaning Steps Performed:

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

	Quote Number	End Customer	Product Hierarchy	Product Family	Geographic Region	Sold-to party	Distributor	Part Number	Quarter of Year	Month of Year	Net Inv QTY	Net Inv-$
0	405011005	TTI ELECTRONICS ASIA PTE LTD.	PPF650NA09 ZZ	Midgit KLKD	ASIA	405011	TTI ELECTRONICS ASIA PTE LTD	KLKD.500T	20174	Period 11 2017	10.0	110.60

- 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	Quarter	Month	Net Inv QTY	Net Inv-$
0	TTI ELECTRONICS ASIA PTE LTD.	PPF650NA09 ZZ	Midgit KLKD	ASIA	405011	TTI ELECTRONICS ASIA PTE LTD	KLKD.500T	Q4	2017-11-01	10.0	110.6
1	TTI ELECTRONICS ASIA PTE LTD.	PPF650NA09 ZZ	Midgit KLKD	ASIA	405011	TTI ELECTRONICS ASIA PTE LTD	KLKD.500T	Q4	2017-12-01	10.0	110.6
2	TTI ELECTRONICS ASIA PTE LTD.	PPF650NA09 ZZ	Midgit KLKD	ASIA	405011	TTI ELECTRONICS ASIA PTE LTD	KLKD.500T	Q1	2018-01-01	10.0	110.6
3	TTI ELECTRONICS ASIA PTE LTD.	PPF650NA09 ZZ	Midgit KLKD	ASIA	405011	TTI ELECTRONICS ASIA PTE LTD	KLKD.500T	Q2	2018-04-01	10.0	110.6

    -

- Missing data:
    - Ran the .info() attribute and realized the “Net Inv Qty column had some missing data.

```
1  # CHECKING THE DIFFERENT COLUMN ENTRIES.
2  solsales.info()
```

```
<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
```

- There are 3764 missing values on the Qty column

```
solsales['Net Inv Qty'].value_counts(dropna = False).head()
```

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

- Looking at all the rows with the NaN values, you quickly realize the associated Net-Inv $ are all negative. This has to do with what is known as “Ship and Debits” from distributors which are not relevant to our analysis

```
solsales[solsales['Net Inv Qty'].isnull()].head(3)
```

	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
2	*PLATT VARIOUS CONTRACTORS	PPF650NA09 ZZ	Midgit KLKD	US/CAN	406311	PLATT ELECTRIC SUPPLY-ROSEVILLE	KLKD015.T	1	2014	Q1	2014-01-01 00:00:00	NaN	-18.90
3	*ARCELOR MITTAL DOFASCO STEEL	PPF650NA09 ZZ	Midgit KLKD	US/CAN	323645	NATIONAL FUSE PRODUCTS	KLKD.500T	1	2014	Q1	2014-01-01 00:00:00	NaN	-97.20

- We apply the dropna() method to drop all missing value rows.

```
# Checking that all the NaN records have been dropped.
slr_sls = solsales.dropna()

slr_sls.info()

slr_sls['Net Inv Qty'].value_counts(dropna = False).head()
```

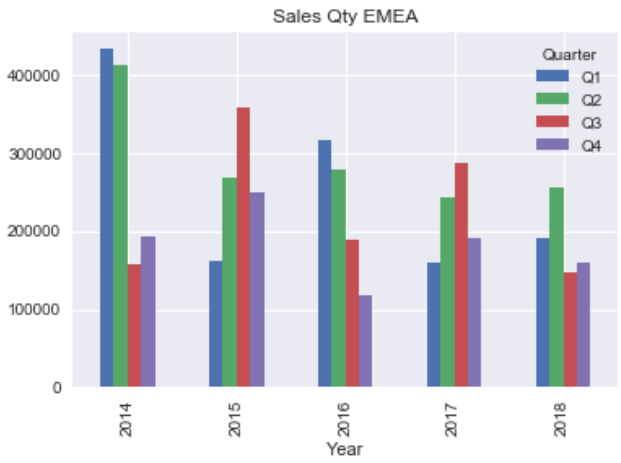
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 26152 entries, 1 to 29915
Data columns (total 13 columns):
End Customer          26152 non-null object
Product Hierarchy     26152 non-null object
Product Family        26152 non-null object
Geographic Region     26152 non-null object
Sold-to party         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
20.0     2377
100.0    2101
30.0     1163
50.0     1134
Name: Net Inv Qty, dtype: int64
```

-

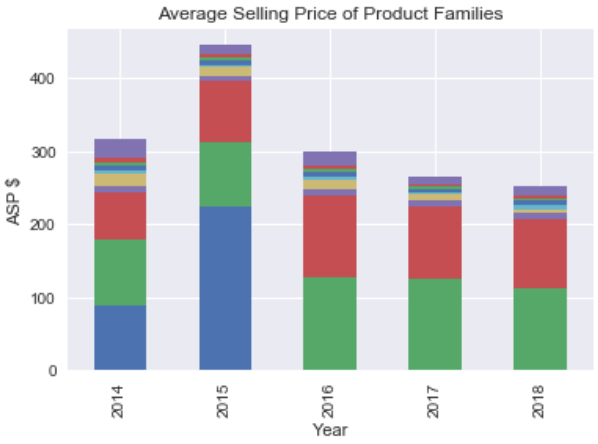
Data Story Telling:

Regional Comparisons by Quarter:

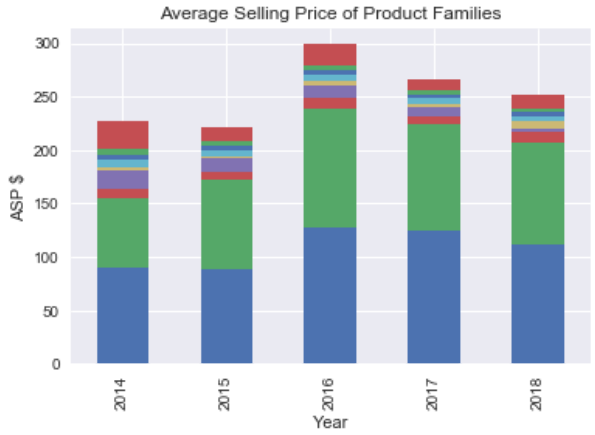
- The strongest markets are the US and Asia followed by Europe and then Latin America
- 2015 was a great year for both US and Asia markets but 2016 saw a steep decline in these markets.
- Since 2014 Europe has been on a consistent decline in units sold while Latin America has been growing slowly having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



- The average selling prices overall have fluctuated but have seen a drop over the years.
- One of the product families “1000Vdc Solar” did not sell in 2016 - 2018 . It contributed to the high ASP total in 2015.



- Taking the this product family off and 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.

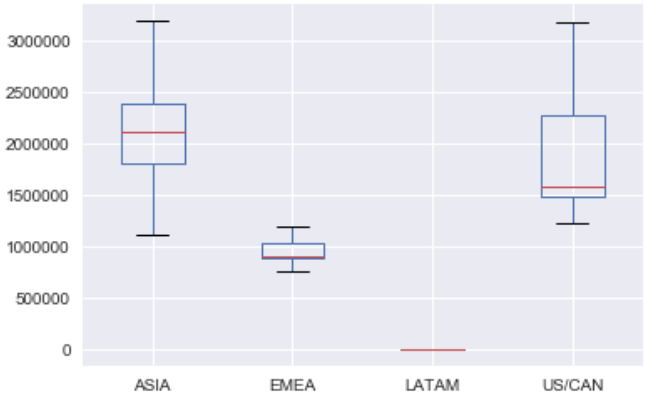


Exploratory Data Analysis

The dataset we set out to analyze is four years (2014 - 2018) of global sales data by month, by quarter and by region of all SKUs sold into the renewable energy industry. 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.

The main variables to consider in the dataset is the quantity of SKUs sold and their year over year variability as shown below. A time series provides of the data provides a visualization of the data:

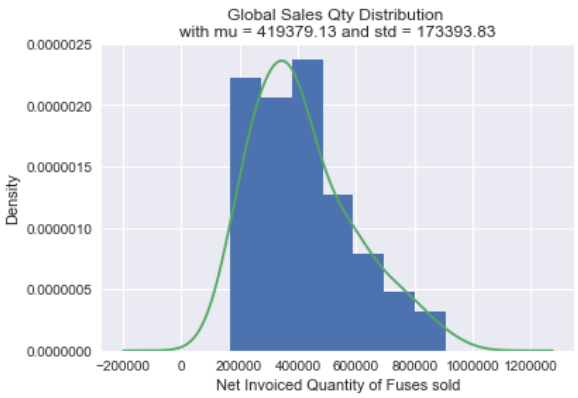
	Net Inv Qty				
Year	2014	2015	2016	2017	2018
count	4.000000e+00	4.000000e+00	4.000000e+00	4.000000e+00	4.000000e+00
mean	8.824770e+05	1.582397e+06	1.070201e+06	1.188924e+06	1.566689e+06
std	5.887321e+05	1.369326e+06	8.072041e+05	1.002952e+06	1.426316e+06
min	2.512000e+03	2.460000e+03	3.296000e+03	6.870000e+03	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

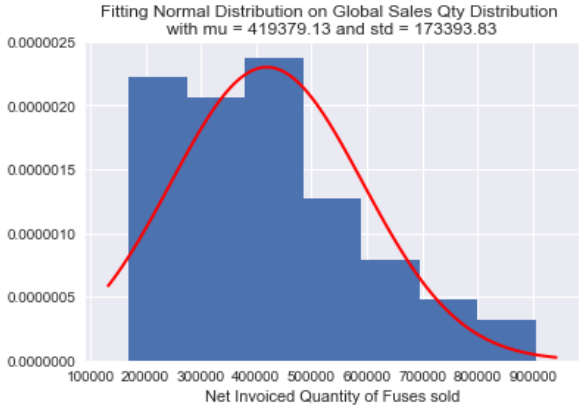


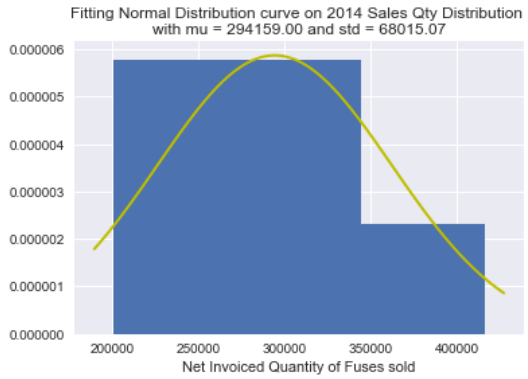
Clearly looking at the take and figure it is clear that the best years so far have been 2015 and 2018. The most variability in terms of quantity sold across the 4 regions is in 2018. The biggest markets have been Asia and the US.

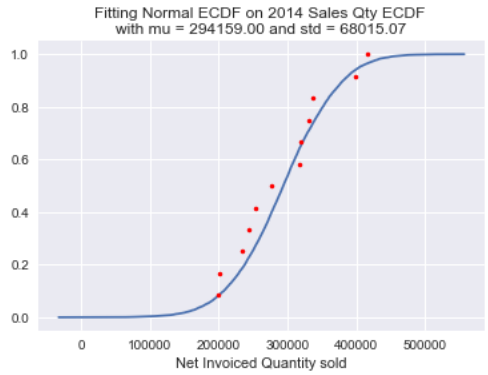
Global Quantity Sales Distribution:

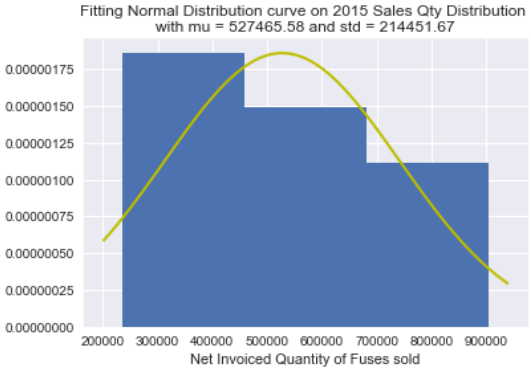
Doing a normality test and plotting the distribution of the total quantity sold globally for the 5 years of data, it gives a right skewed distribution. Which indicates that there are fewer months in which the very high quantities were bought.

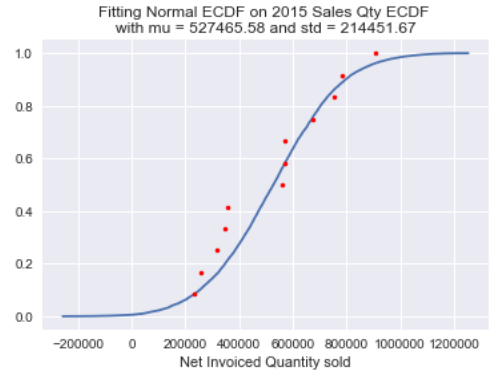


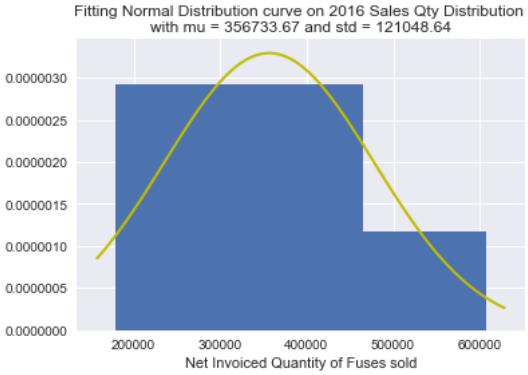


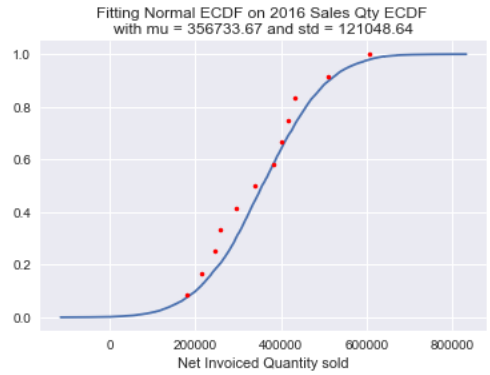

Fitting Normal Distribution curve on 2014 Sales Qty Distribution
with mu = 294159.00 and std = 68015.07
0.000006
0.000005
0.000004
0.000003
0.000002
0.000001
0.000000
200000
250000
300000
350000
400000
Net Invoiced Quantity of Fuses sold

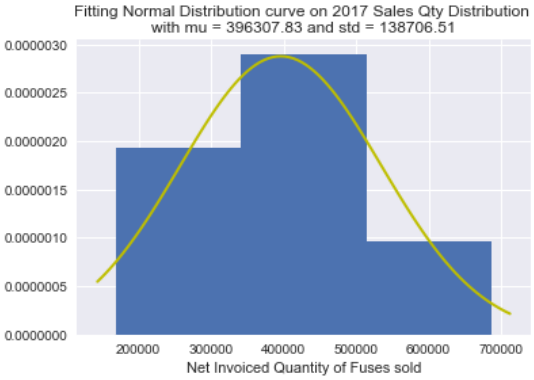

Fitting Normal ECDF on 2014 Sales Qty ECDF
with mu = 294159.00 and std = 68015.07
1.0
0.8
0.6
0.4
0.2
0.0
0
100000
200000
300000
400000
500000
Net Invoiced Quantity sold

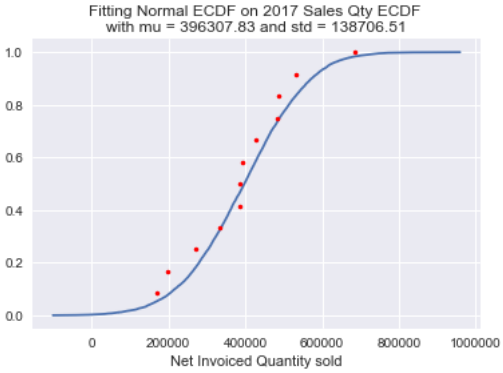

Fitting Normal Distribution curve on 2015 Sales Qty Distribution
with mu = 527465.58 and std = 214451.67
0.00000175
0.00000150
0.00000125
0.00000100
0.00000075
0.00000050
0.00000025
0.00000000
200000
300000
400000
500000
600000
700000
800000
900000
Net Invoiced Quantity of Fuses sold

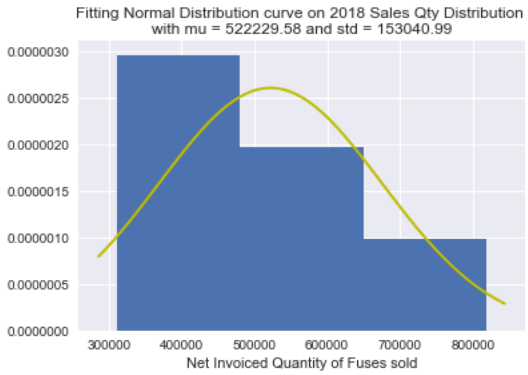

Fitting Normal ECDF on 2015 Sales Qty ECDF
with mu = 527465.58 and std = 214451.67
1.0
0.8
0.6
0.4
0.2
0.0
−200000
0
200000
400000
600000
800000
1000000
1200000
Net Invoiced Quantity sold

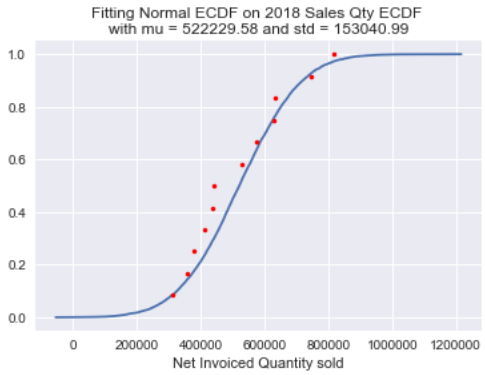

Fitting Normal Distribution curve on 2016 Sales Qty Distribution
with mu = 356733.67 and std = 121048.64
0.0000030
0.0000025
0.0000020
0.0000015
0.0000010
0.0000005
0.0000000
200000
300000
400000
500000
600000
Net Invoiced Quantity of Fuses sold


Fitting Normal ECDF on 2016 Sales Qty ECDF
with mu = 356733.67 and std = 121048.64
1.0
0.8
0.6
0.4
0.2
0.0
0
200000
400000
600000
800000
Net Invoiced Quantity sold


Fitting Normal Distribution curve on 2017 Sales Qty Distribution
with mu = 396307.83 and std = 138706.51
0.0000030
0.0000025
0.0000020
0.0000015
0.0000010
0.0000005
0.0000000
200000
300000
400000
500000
600000
700000
Net Invoiced Quantity of Fuses sold


Fitting Normal ECDF on 2017 Sales Qty ECDF
with mu = 396307.83 and std = 138706.51
1.0
0.8
0.6
0.4
0.2
0.0
0
200000
400000
600000
800000
1000000
Net Invoiced Quantity sold





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