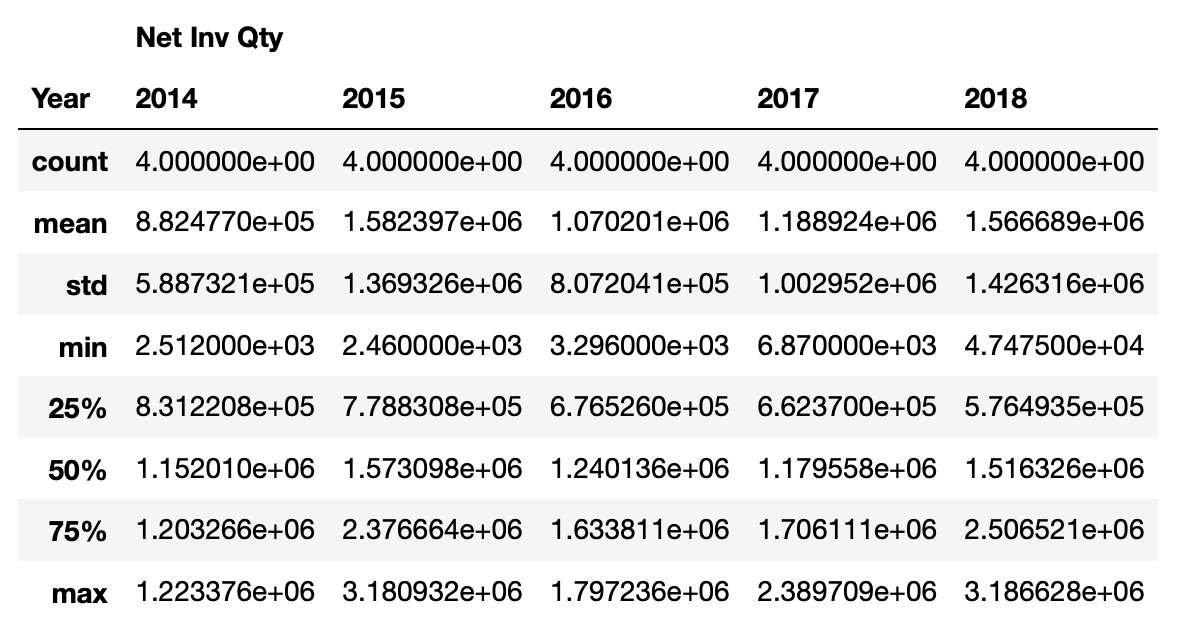
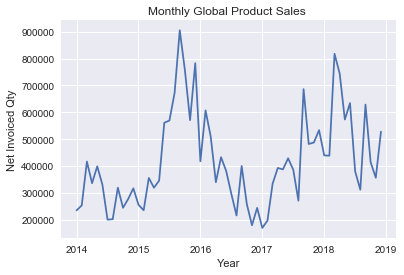
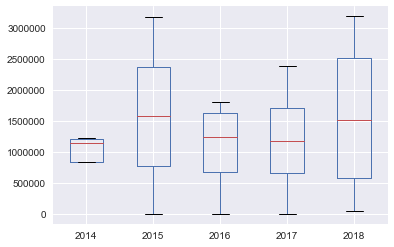
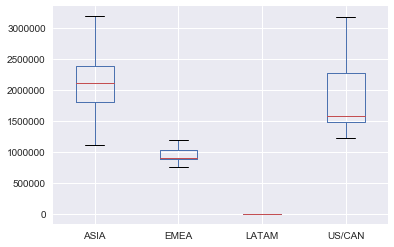
**Capstone Project 1: Exploratory Data Analysis**

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:

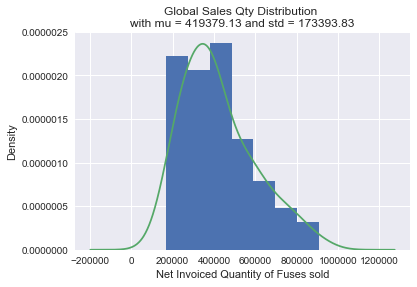
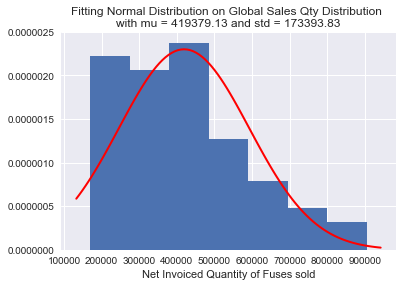
 

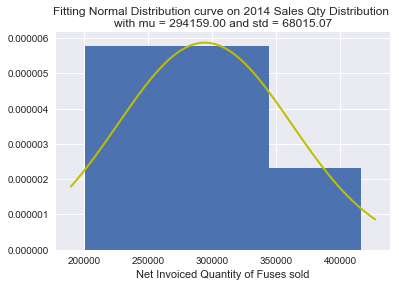
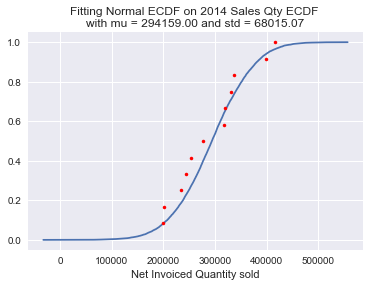
 

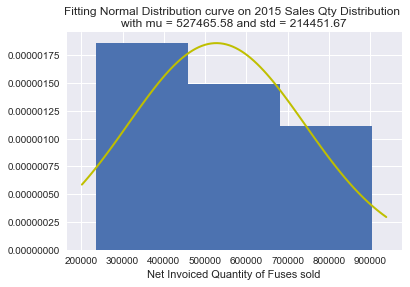
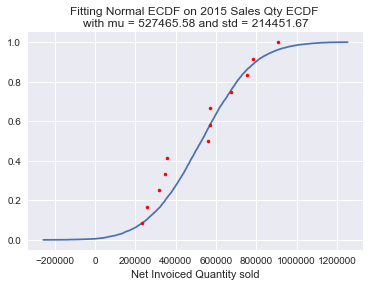
Clearly looking at the take and figure it is clear that the best years so far have been 2015 and 2018. The most variability in terms of quantity sold across the 4 regions is in 2018. The biggest markets have been Asia and the US.

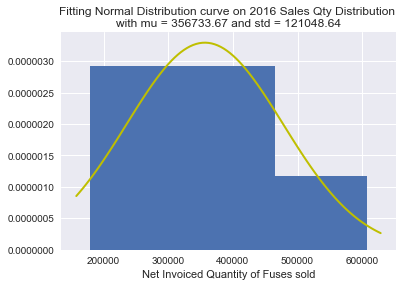
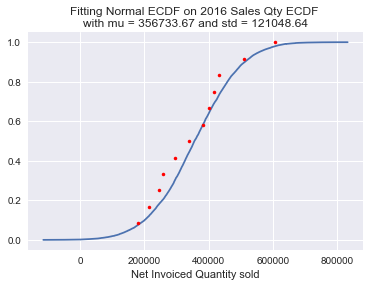
**Global Quantity Sales Distribution:**

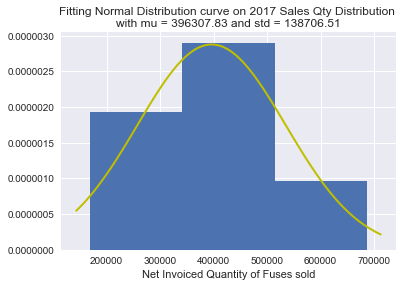
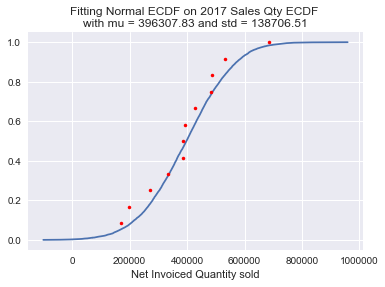
Doing a normality test and plotting the distribution of the total quantity sold globally for the 5 years of data, it gives a right skewed distribution. Which indicates that there are fewer months in which the very high quantities were bought.

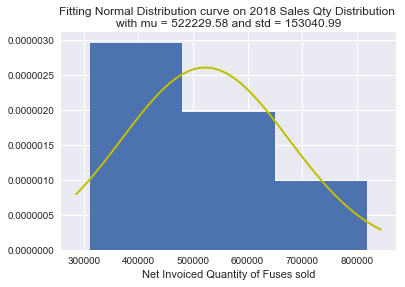
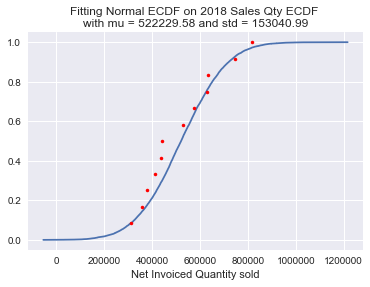
 

The data varies greatly from year to year and there does not seem to be a specific trend in the sales. To better forecast the sales quantities, we will need to account for the variability and seasonality of the data. One of the ways to do this is to use an ARIMA model.