Out[1]: The raw code for this IPython notebook is by default hidden for easier reading. To toggle on/off the raw code, click here

Story Telling

This book is to assist on presenting the findings inside a data sample provided. For the courious minds, you can see what is behind the scenes by reviewing the main notebook on this directory (main.ipynb)

How to use it?

You can go cells by cell reading the content and associating the readings to the charts presented. Altough you can skip and go to a particular it's recommended to go in the sequencial order presented to acquire the overal picture. then you can go back and forth as you wish.

Can this book be reused?

You can provide a new dataset and follow the steps below to obtain the view of the new dataset:

- Load the file with the new data on the same directory as this file.
- Open the `main.ipynb` book and run all the cells in it.
  - a new set of charts will be generated
  - a set of test will be ran on the data.
- Close this book and re-open it.
- When reopened, this book will refresh with the new charts.
- The comments on this book are applicable to the current dataset, for the new dataset the comments need to be revisited to validate them against the new charts and the new tests made. (in other words, the charts are automatically generated but the comments need to be reviewed for new dataset)
- Save this file to keep the comments made based on the new dataset.

What is the problem at hand?

A given set of data is provided and the goals are:

- to determine the type of data and how is its broken down
- identify if the data is normally distributed
- forecast the expected values until the end of 2021.

Let's start... on the top menu select `Kernel` and then `Restart and Run All Cells`

Understanding the data

Methodology: Discover the contents of the information provied and using python and packages like pandas, sklearn, seaborn and others identify the best way to predict the future values.

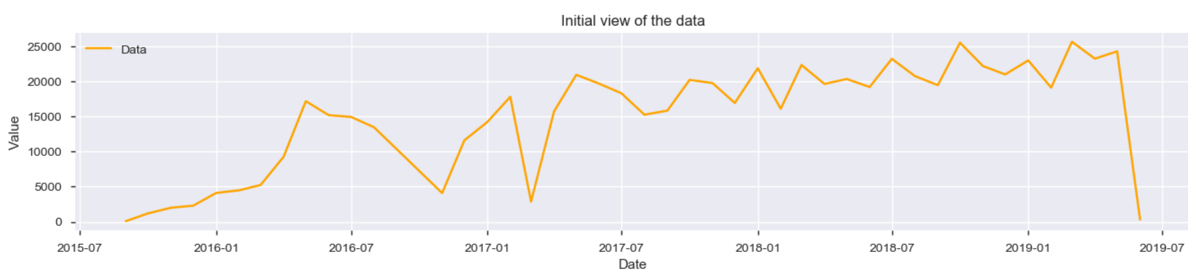
```
After consolidation we have 515 records with the following structure:
          date Status Branch  Value
0   2015-09-01      C    FR1     70
513 2019-06-01      C    FR1    310
```

Data contents

```
Summary of the data contents:
        Status codes (3):       'C', 'UE', 'EE'
        Branch names (5):       'FR1', 'IT2', 'SP3', 'GE4', 'HK5':
        Start date:             2015-09-01 00:00:00
        End date:               2019-06-01 00:00:00
        Months in the data:     46
        Expected records:       690
        Consolidated records:   515
```

How does the data look like?

```
Initial view of the data
```



- Clearly the first and last points in the chart are outliers with some others in the middle
- Following are some of the potential outliers

Out[7]:

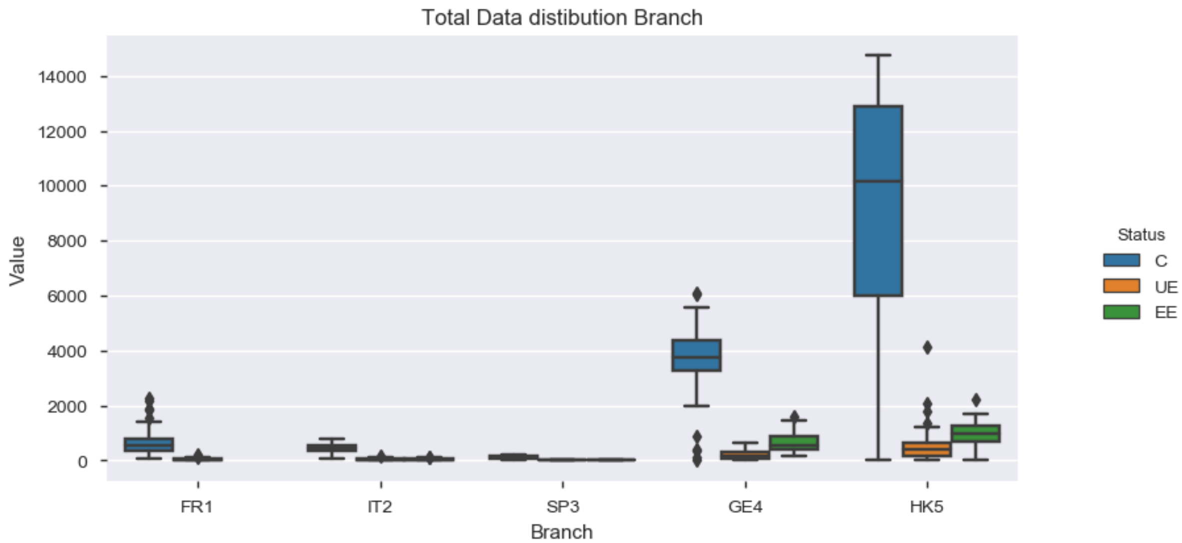
	Smallest values	Largest values
date		
2015-09-01	85	
2015-10-01	1196	
2015-11-01	2000	
2015-12-01	2296	
2019-06-01	315	
2018-07-01		23237
2018-10-01		25534
2019-03-01		25668
2019-04-01		23256
2019-05-01		24296

Distribution of observations among categorical values

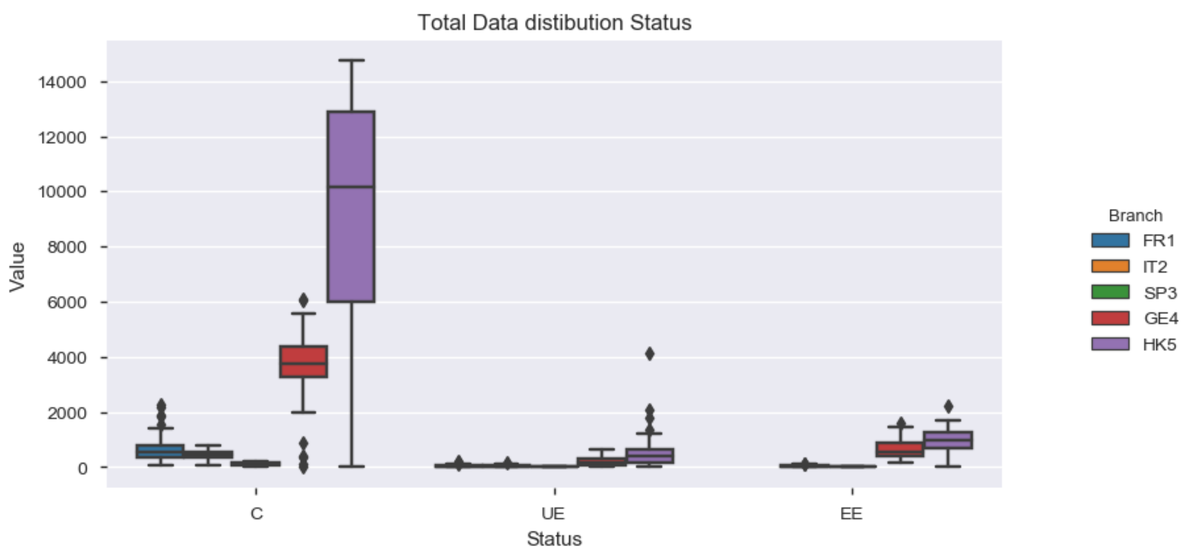
Out[8]:

	Count C	Count EE	Count UE	Sum C	Sum EE	Sum UE	sum	% C	% EE	% UE
Branch										
FR1	44		44							
GE4	41	34	41							
HK5	41	41	41							
IT2	40	29	38							
SP3	41	2	38							
FR1				30950	0	2288	33238	4.62%	0.00%	0.34%
GE4				147511	22827	8090	178428	22.00%	3.41%	1.21%
HK5				369068	38062	25091	432221	55.05%	5.68%	3.74%
IT2				18365	1458	1527	21350	2.74%	0.22%	0.23%
SP3				4688	3	454	5145	0.70%	0.00%	0.07%

- There is almost the same amount of records per branch in the Status `C` and `UE` but that is not the case for status `EE` as SP3 and FR1 are practically null.
- For status `C` : The sum of the values for SP3 represent a very low percentage of the total. The sum of the values for GE4 and HK5 sum of values represents the 77% which will drive the overall results. This is more evident in the charts below

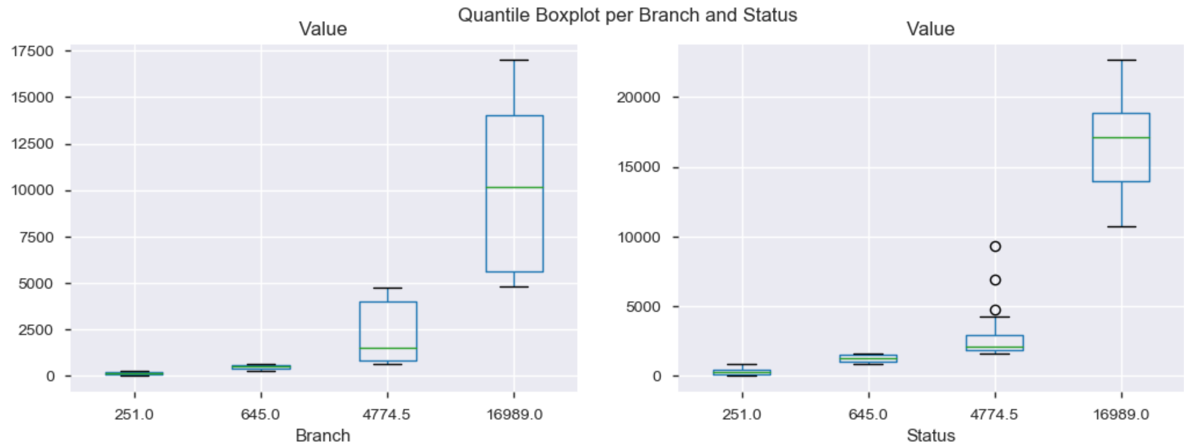


- We can see that HK5 has the sample with the larger values. FR1, IT2 and SP3 look really small when compared to it. #### Distribution by Status



- We can see that HK5 has the sample with extreme values for Status = C. For Status UE again HK5 is the main driver and has clear outliers.

Quantile distribution



- Branch values are heavily accumlated in the 3rd and 4th quantiles
- Status values are accumulated in the 4th quantile and 3rd quantile has several outliers

Test for Normal Distribution of the data

Methodology:

Building an Empirical Cumulative Distribution Function (**ECDF**) to visualize the distribution of the data and using **scipy.stats.normaltest** test whether the sample differs from a normal distribution.

This function tests the **`null hypothesis that a sample comes from a normal distribution`** . It is based on D'Agostino and Pearson's test that combines skew and kurtosis to produce an omnibus test of normality.

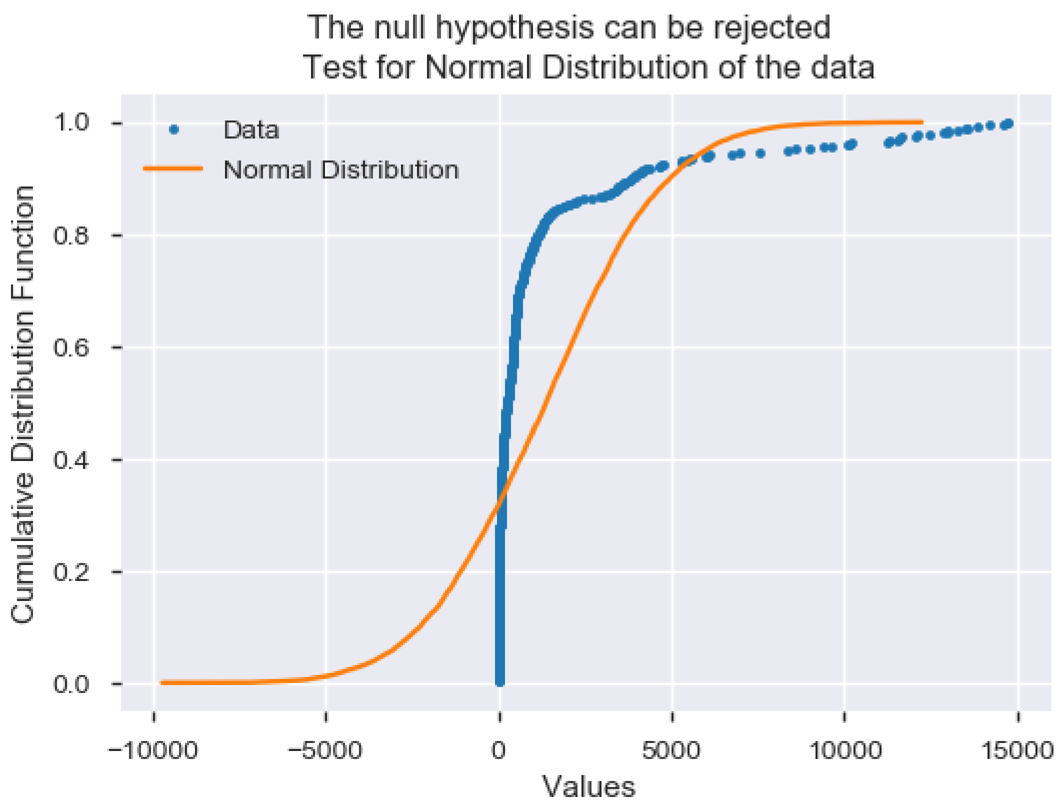
The `p value` shown will validate wheter or not the data is normally distributed.

We can adjust aplha to indicate to which percentage of precision we want to validate the data normallity. on this case we will use 0.1 meaning if the data is within a 10% of the normal distribution, the data will be considered normally distributed this is the **`Null hypothesis`** .

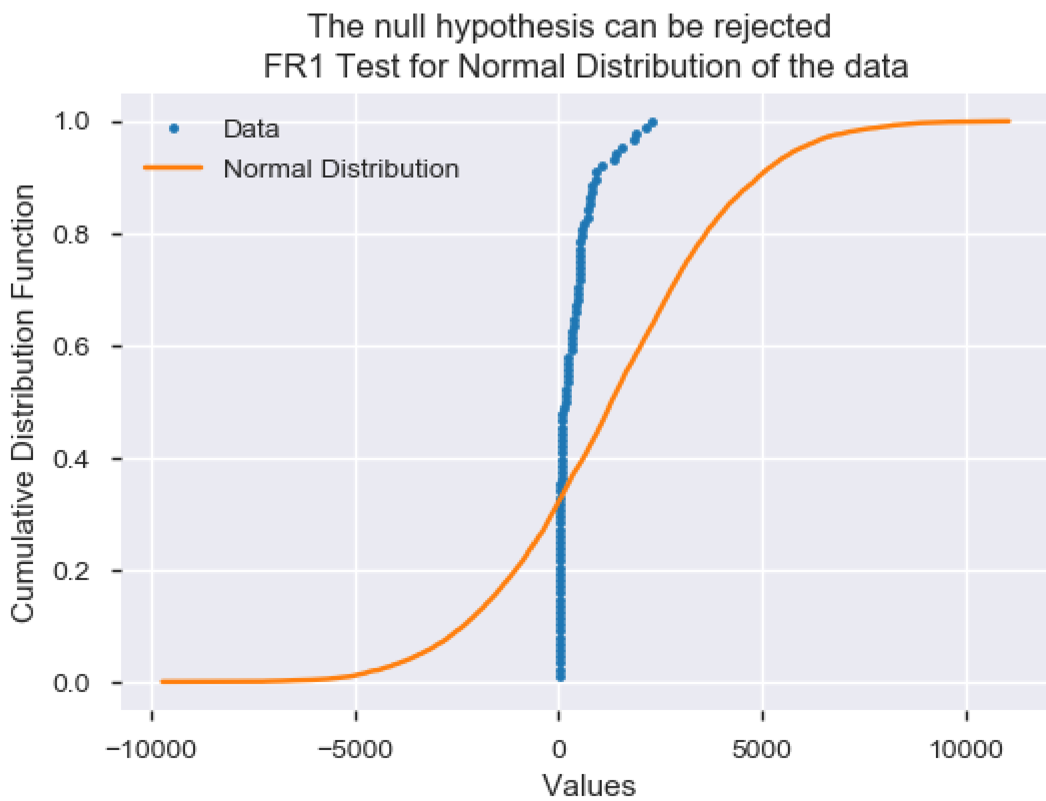
If `p value` is greater than alpha we cannot reject the null hypothesis and must conclude the data is normally distributed.

Testing on the monthly series

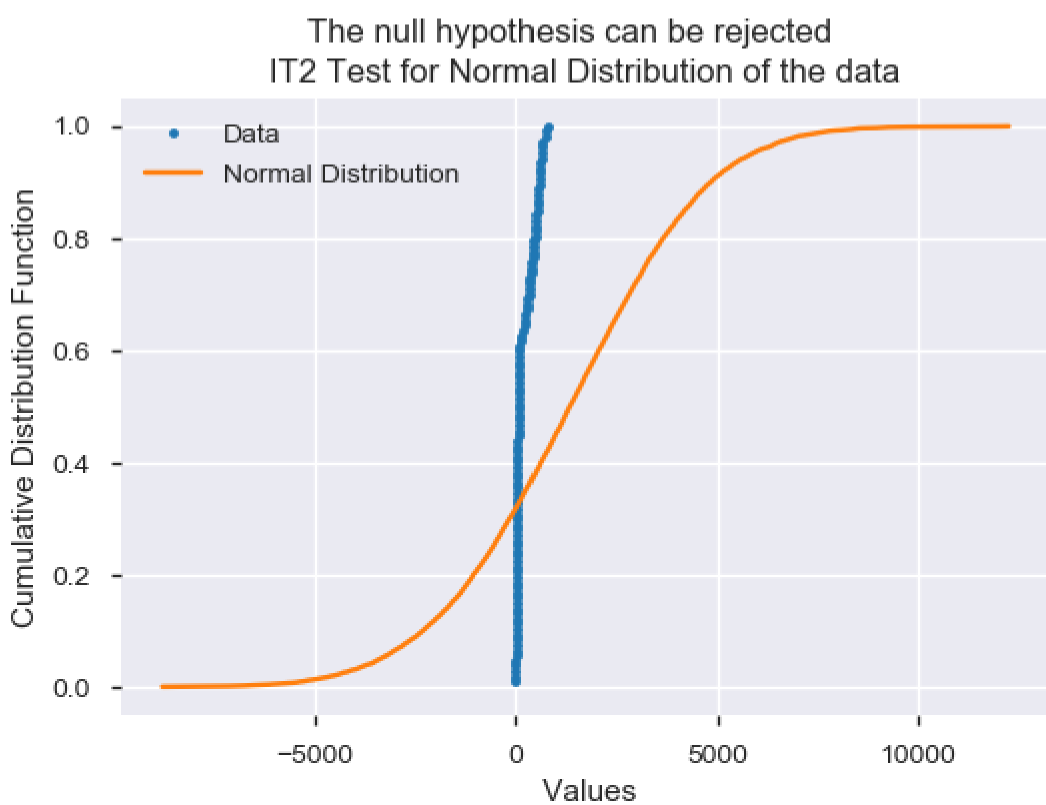
```
(s^2 + k^2): 358.2853737070243  pvalue: 1.582412213227274e-78
p = 1.58241e-78 alplha = 0.100
The null hypothesis can be rejected
```



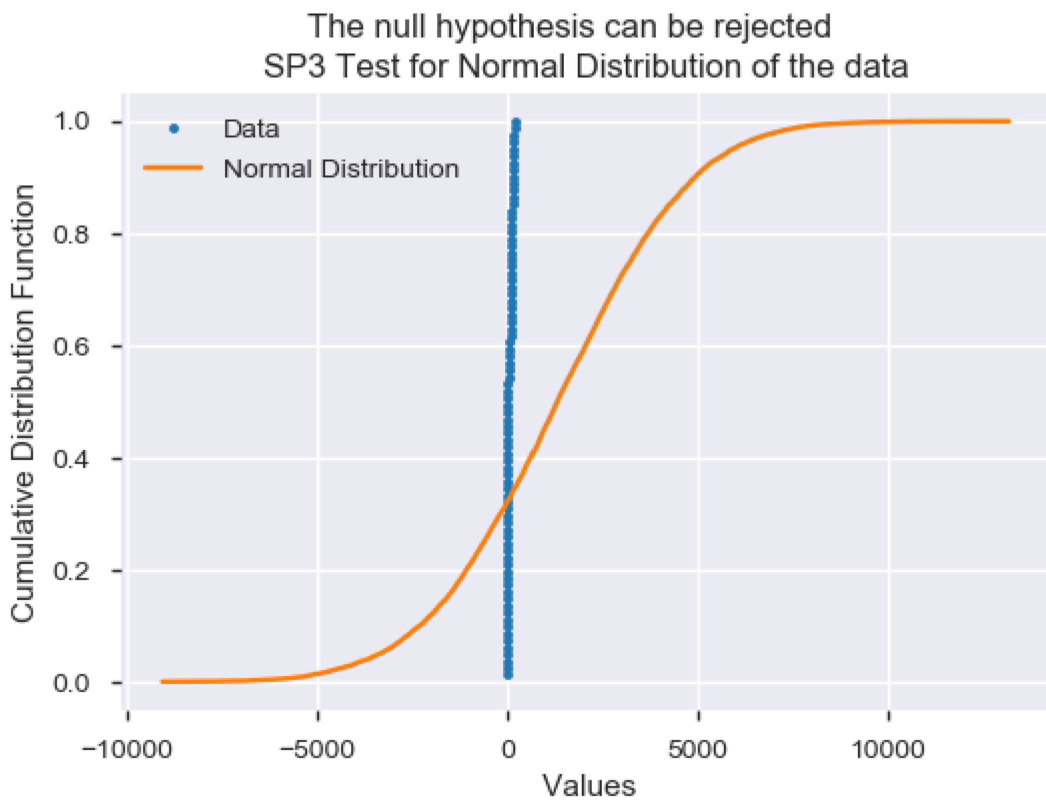

```
(s^2 + k^2): 47.754208968062564 pvalue: 4.2687947989147863e-11
p = 4.26879e-11 alplha = 0.100
The null hypothesis can be rejected
```



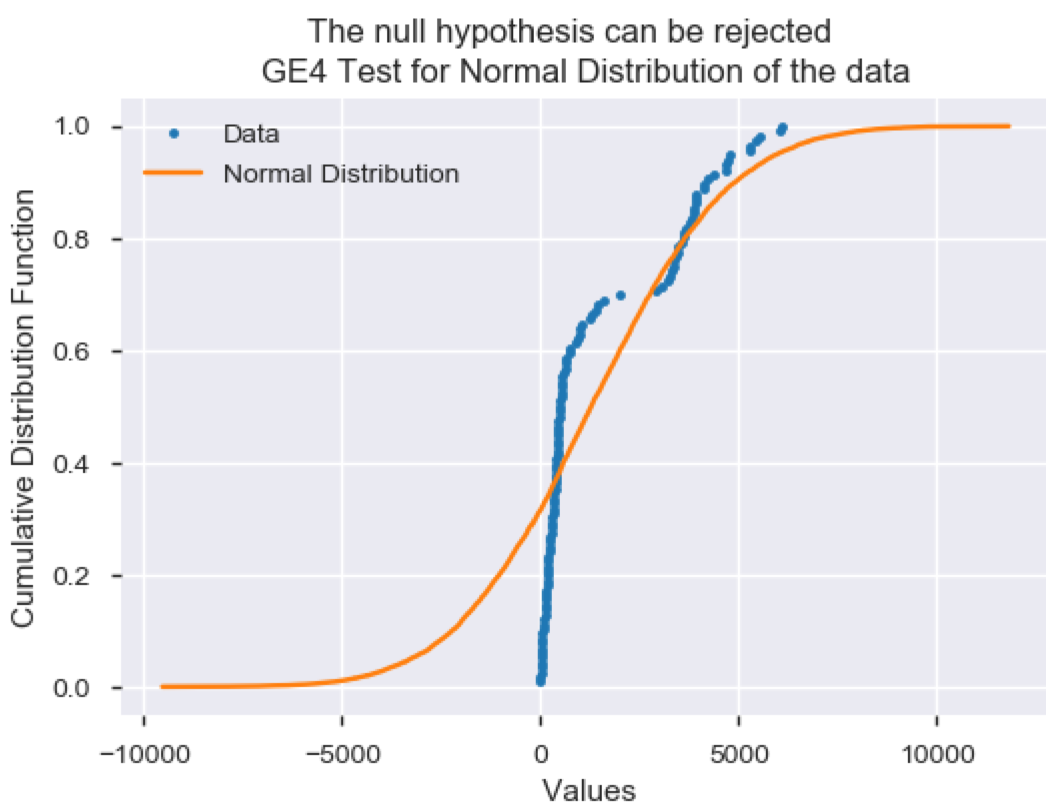

```
(s^2 + k^2): 15.741421490720878 pvalue: 0.0003817629295207696
p = 0.000381763 alplha = 0.100
The null hypothesis can be rejected
```



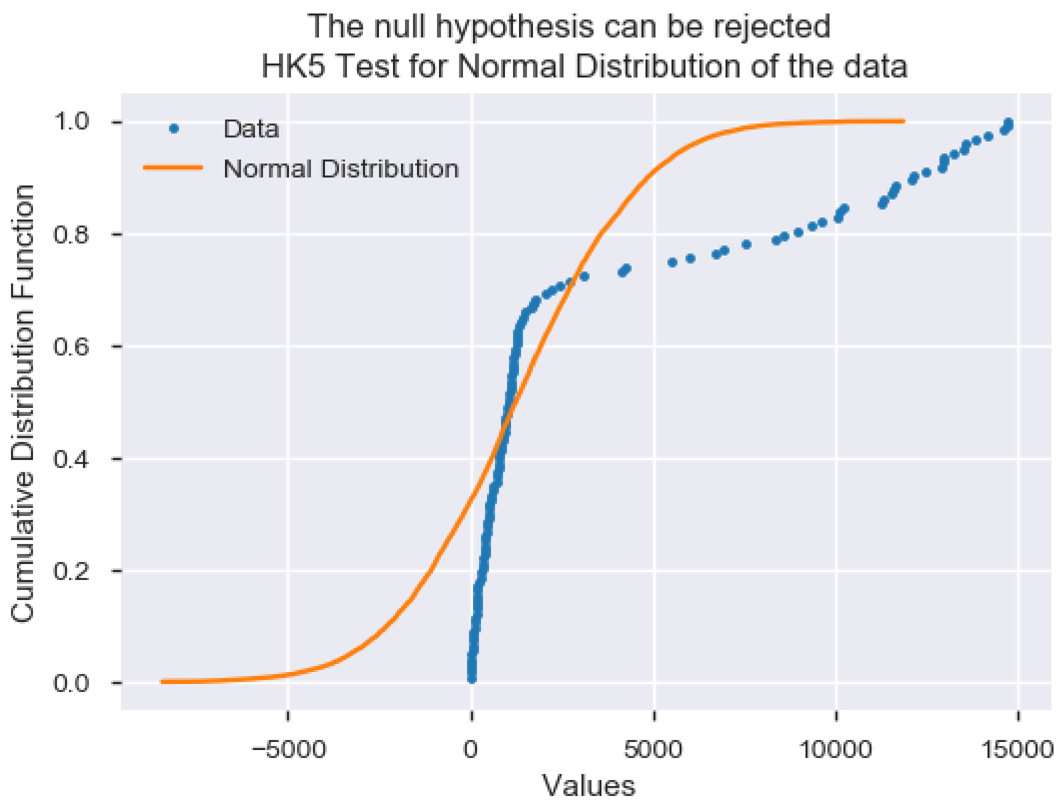

```
(s^2 + k^2): 10.672361239506003 pvalue: 0.0048142229890101515
p = 0.00481422  alplha = 0.100
The null hypothesis can be rejected
```




```
(s^2 + k^2): 17.49566986827648  pvalue: 0.00015880477598388283
p = 0.000158805 alplha = 0.100
The null hypothesis can be rejected
```

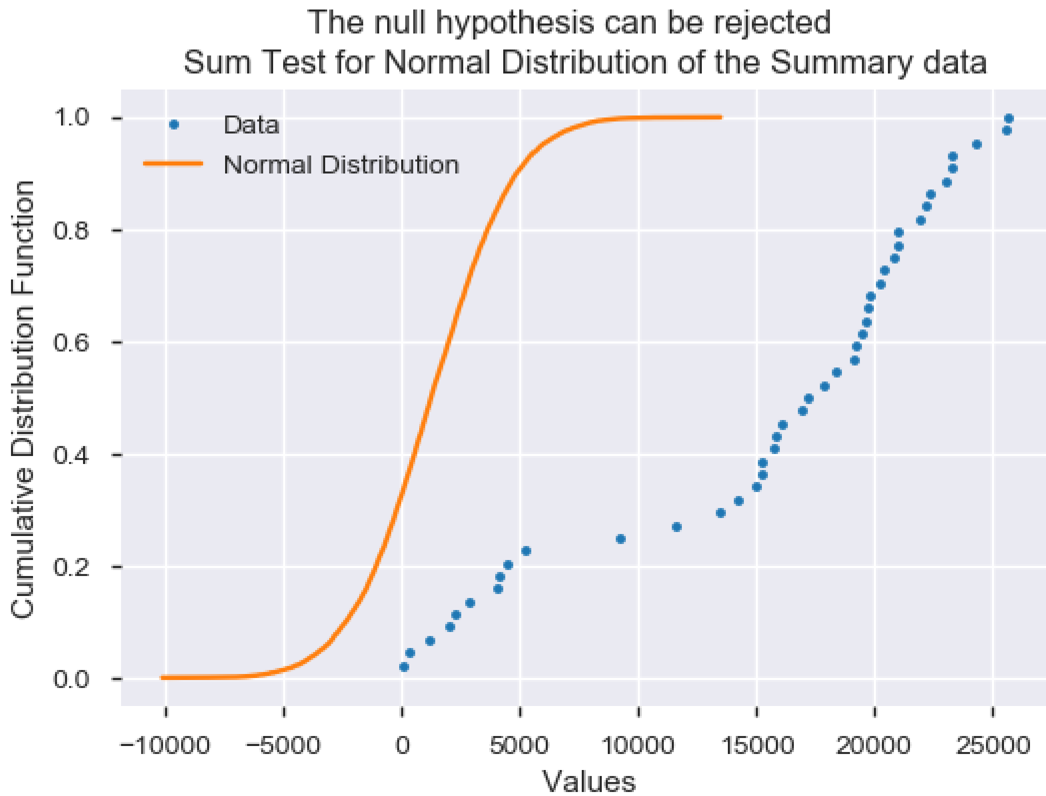



```
(s^2 + k^2): 23.92121707051891  pvalue: 6.391072039992402e-06
p = 6.39107e-06 alplha = 0.100
The null hypothesis can be rejected
```

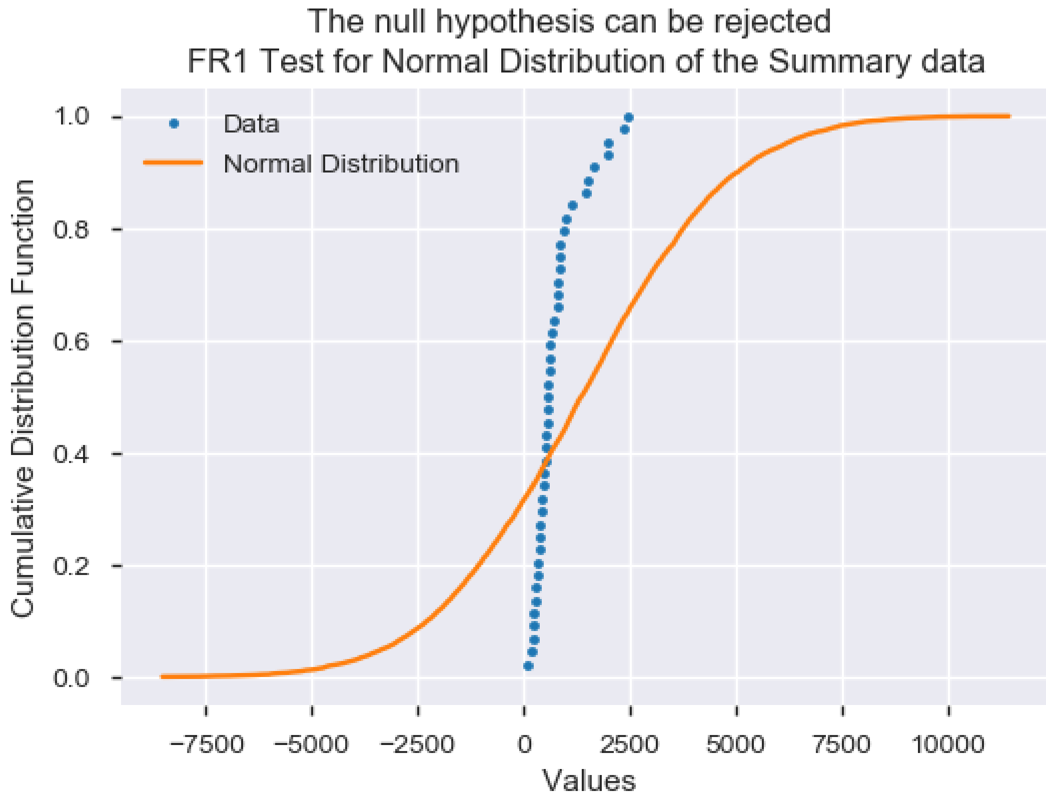


Testing on the summary series (Status values consolidated per branch and the total data)

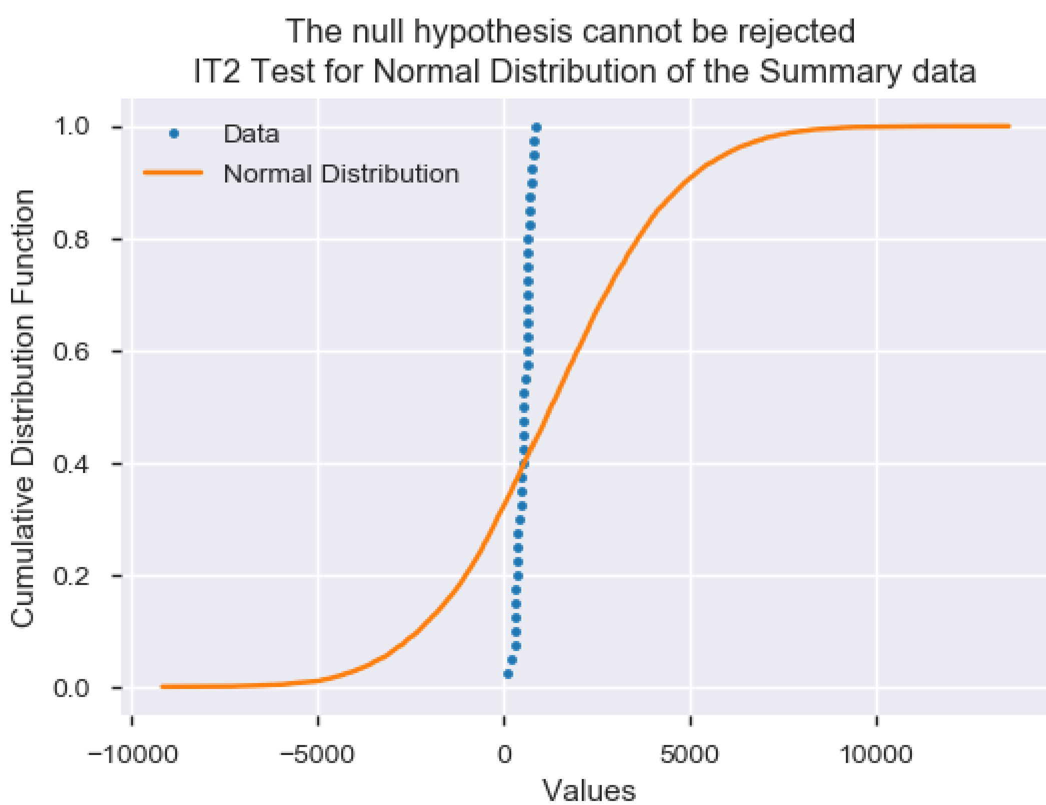
```
(s^2 + k^2): 5.961616636167258  pvalue: 0.05075179368775998
p = 0.0507518   alplha = 0.100
The null hypothesis can be rejected
```



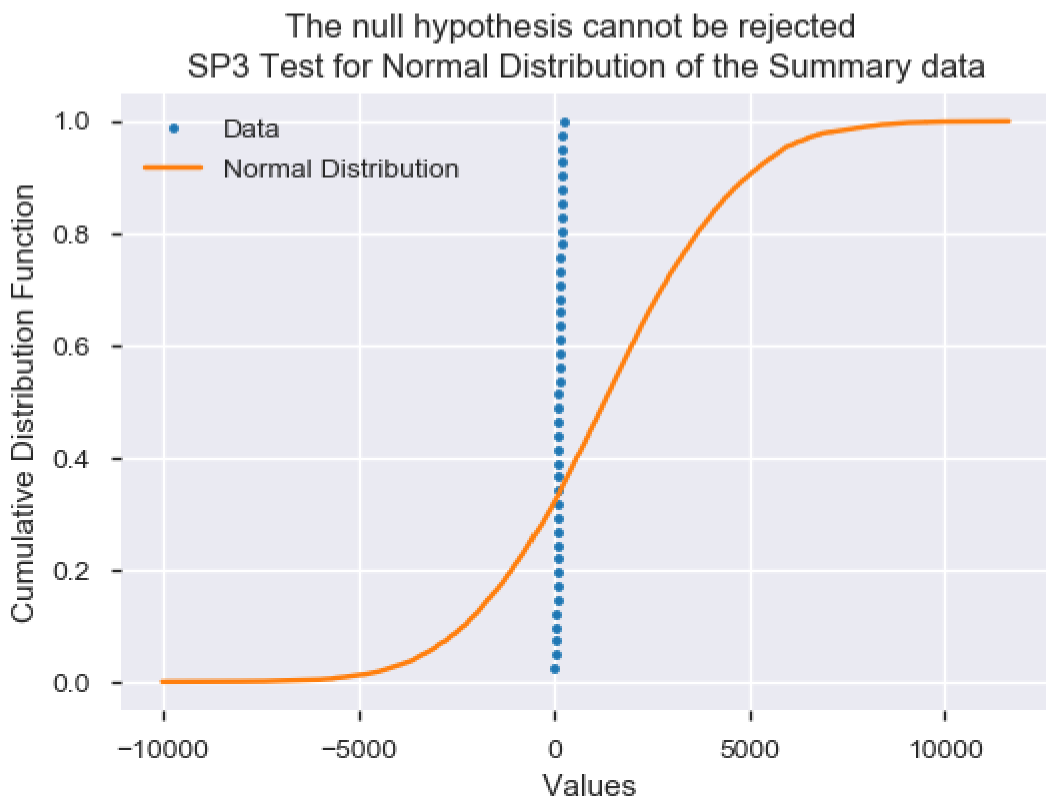

```
(s^2 + k^2): 18.1632000184431   pvalue: 0.00011373947757791924
p = 0.000113739 alplha = 0.100
The null hypothesis can be rejected
```



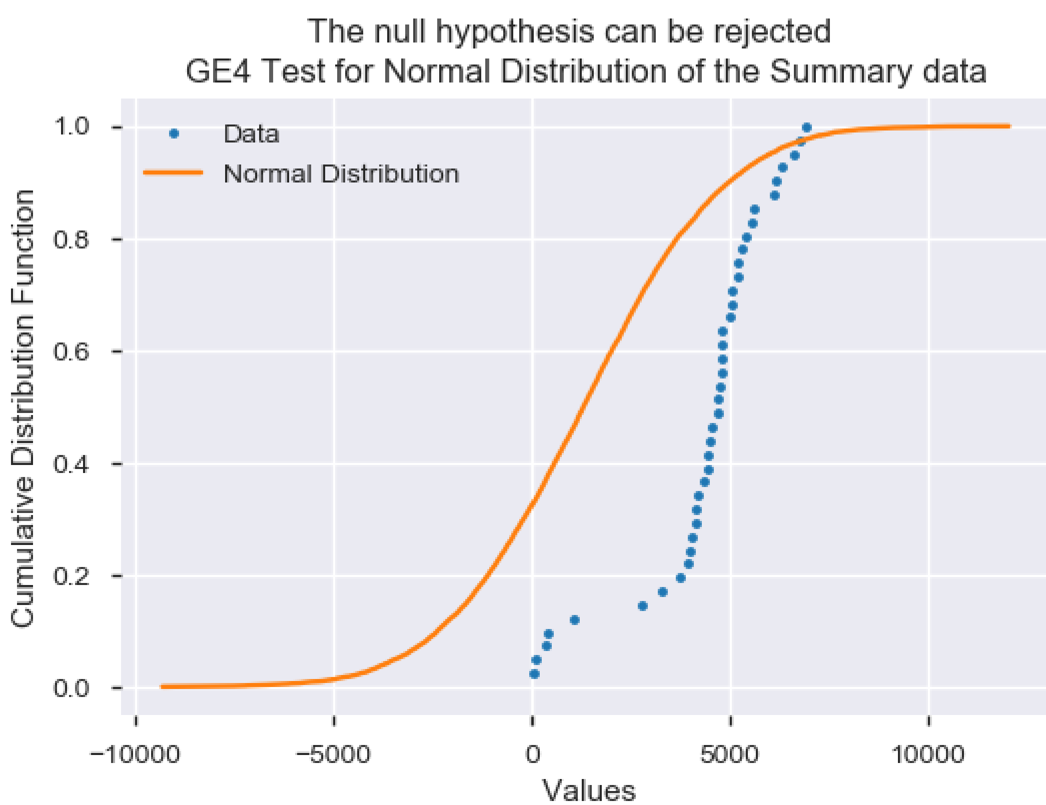

```
(s^2 + k^2): 1.2700405560167773 pvalue: 0.5299247423902442
p = 0.529925    alplha = 0.100
The null hypothesis cannot be rejected
```



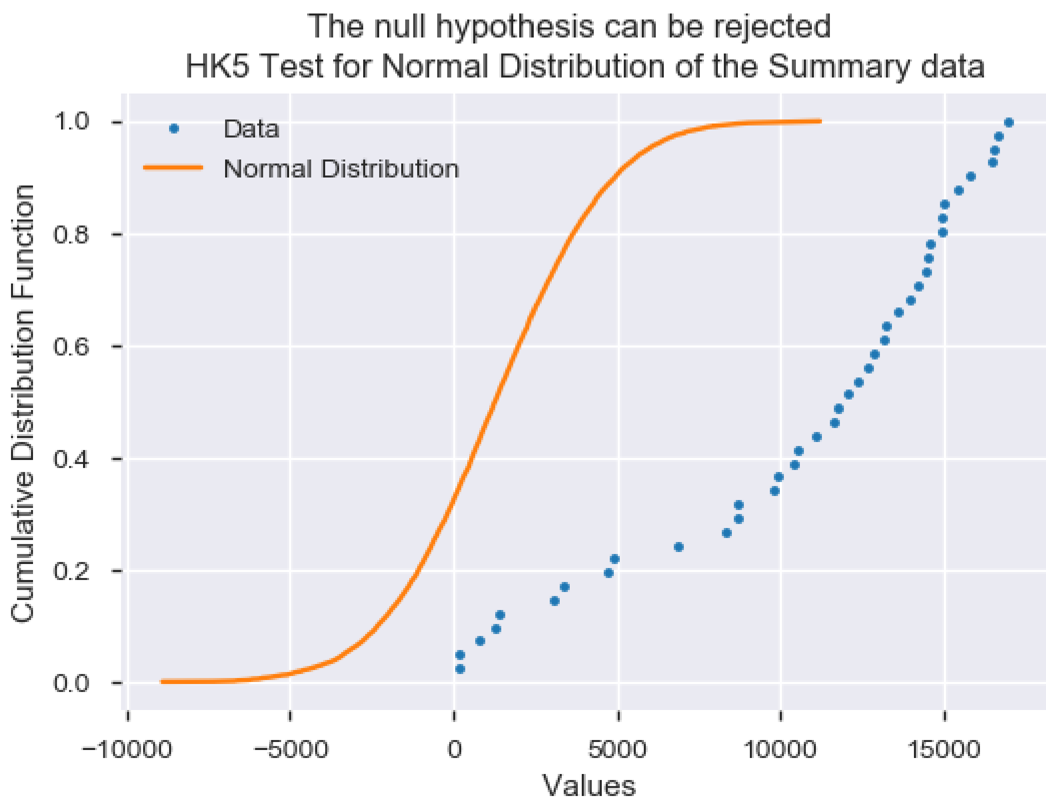

```
(s^2 + k^2): 0.032095922139047654       pvalue: 0.984080121388326
p = 0.98408     alplha = 0.100
The null hypothesis cannot be rejected
```




```
(s^2 + k^2): 12.092873679947456 pvalue: 0.0023662784317696325
p = 0.00236628  alplha = 0.100
The null hypothesis can be rejected
```



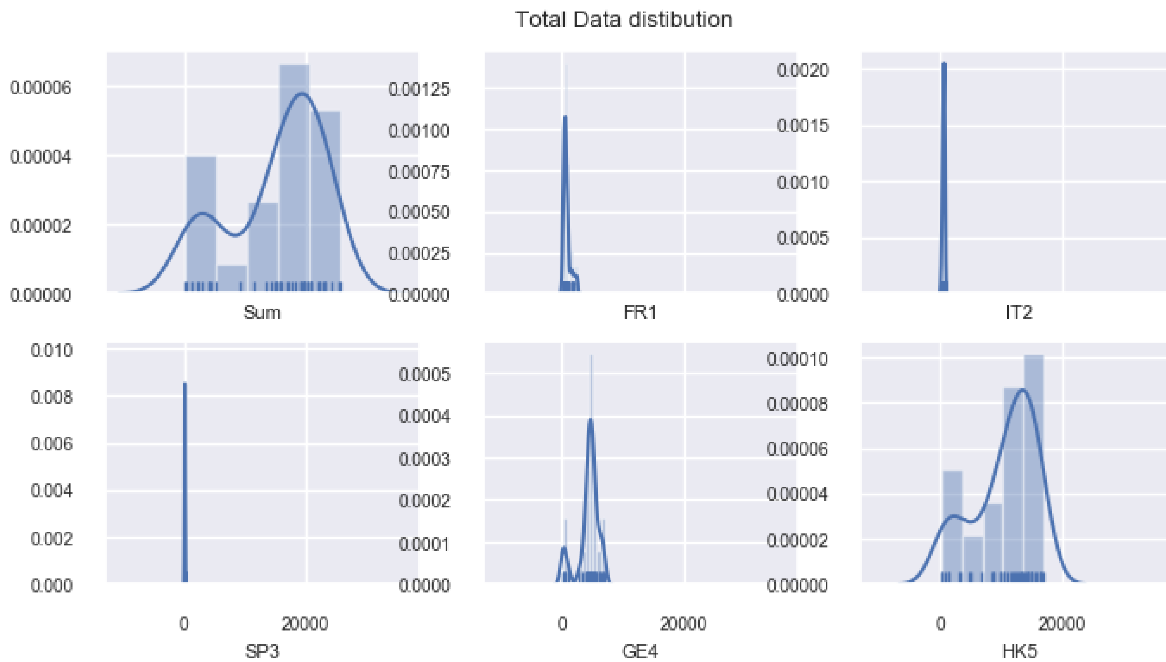

```
(s^2 + k^2): 5.103102139692429  pvalue: 0.07796064976005972
p = 0.0779606   alplha = 0.100
The null hypothesis can be rejected
```



- as the test result indicates, "The null hypothesis can be rejected" we conclude the data is not normally distributed.
- Similar methodology was applied to all the branch in both the Monthly and consolidated data. results shown on the charts indicate that none of them is normally distribuited within a 10% margin (10% is a big range but can be adjusted as needed)

Descriptive Statistics

```
Total Data distibution :
```



- The continuous line is trying to fit the data to a normal distribution. We can see in summary (Sum) the data has two dome's comparing this with the HK5, we see similarities between them due to the fact that HK5 has the larger values as previously shown.

Descriptive Statistics Monthly Series

These are the values after aggregating the data per Branch. (but Status are NOT aggregated).
The Sum represents the total for all branches (status not consolidated)

Descriptive Statistics Monthly Series

Out[16]:

Description	Total	FR1	IT2	SP3	GE4	HK5
Arithmetic mean ('average') of data	1,301.71	377.70	199.53	63.52	1,538.17	3,513.99
Harmonic mean of data	29.59	57.24	15.72	11.18	94.55	207.83
Median (middle value) of data	273.00	158.00	69.00	29.00	491.00	1,054.00
Low median of data	273.00	153.00	69.00	29.00	484.00	1,054.00
High median of data	273.00	163.00	69.00	29.00	498.00	1,054.00
Median, or 50th percentile, of grouped data	273.00	162.50	69.00	29.25	497.50	1,054.00
Mode (most common value) of discrete data	5.00	153.00	63.00	5.00	484.00	1,042.00
Population standard deviation of data	2,809.44	499.63	221.52	62.09	1,780.01	4,712.93
Population variance of data	7,892,967.42	249,634.96	49,072.70	3,855.06	3,168,420.14	22,211,708.97
Sample standard deviation of data	2,812.17	502.50	222.57	62.48	1,787.73	4,732.21
Sample variance of data	7,908,323.39	252,504.33	49,535.65	3,903.25	3,195,971.62	22,393,772.16

Descriptive Statistics Monthly Series

Out[17]:

	Total	FR1	IT2	SP3	GE4	HK5
count	515.00	88.00	107.00	81.00	116.00	123.00
mean	1,301.71	377.70	199.53	63.52	1,538.17	3,513.99
std	2,812.17	502.50	222.57	62.48	1,787.73	4,732.21
min	1.00	5.00	1.00	1.00	3.00	6.00
25%	49.00	41.00	35.00	7.00	242.00	420.00
50%	273.00	158.00	69.00	29.00	491.00	1,054.00
75%	831.00	529.75	385.00	111.00	3,369.25	5,755.50
max	14,748.00	2,270.00	779.00	221.00	6,087.00	14,748.00

Descriptive Statistics Summary

These are the values after aggregating the data per Branch. (all kind of Status grouped by Branch).
The Sum represents the monthly aggregation (all branches and all status consolidated)

```
Descriptive Statistics Summary
```

Out[18]:

Description	Sum	FR1	IT2	SP3	GE4	HK5
Arithmetic mean ('average') of data	15,235.95	755.41	533.75	125.49	4,351.90	10,541.98
Harmonic mean of data	2,216.91	456.29	455.20	75.02	906.16	2,451.52
Median (middle value) of data	17,506.00	567.00	530.50	115.00	4,720.00	12,075.00
Low median of data	17,192.00	562.00	522.00	115.00	4,720.00	12,075.00
High median of data	17,820.00	572.00	539.00	115.00	4,720.00	12,075.00
Median, or 50th percentile, of grouped data	17,819.50	571.50	538.50	115.00	4,720.00	12,075.00
Mode (most common value) of discrete data	17,192.00	527.00	517.00	113.00	4,701.00	11,763.00
Population standard deviation of data	7,644.59	568.72	170.86	50.54	1,710.66	5,067.44
Population variance of data	58,439,782.63	323,443.24	29,192.94	2,554.30	2,926,368.19	25,678,911.44
Sample standard deviation of data	7,732.97	575.30	173.04	51.17	1,731.91	5,130.39
Sample variance of data	59,798,847.35	330,965.18	29,941.47	2,618.16	2,999,527.39	26,320,884.22

```
Descriptive Statistics Summary Series
```

Out[19]:

	Sum	FR1	IT2	SP3	GE4	HK5
count	44.00	44.00	40.00	41.00	41.00	41.00
mean	15,235.95	755.41	533.75	125.49	4,351.90	10,541.98
std	7,732.97	575.30	173.04	51.17	1,731.91	5,130.39
min	85.00	85.00	122.00	6.00	42.00	174.00
25%	11,027.50	370.75	389.50	100.00	4,031.00	8,341.00
50%	17,506.00	567.00	530.50	115.00	4,720.00	12,075.00
75%	20,823.50	855.75	648.75	161.00	5,215.00	14,519.00
max	25,668.00	2,423.00	856.00	233.00	6,931.00	16,989.00

Out[20]:

Status	**C**					**EE**				**UE**			
Branch	**FR1**	**GE4**	**HK5**	**IT2**	**SP3**	**GE4**	**HK5**	**IT2**	**SP3**	**FR1**	**GE4**	**HK5**	**IT2**
count	45	42	42	41	42	35	42	30	3	45	42	42	3
mean	1375	7024	17574	895	223	1304	1812	97	2	101	385	1194	7
std	4540	22255	55744	2800	707	3762	5751	258	1	335	1226	3845	24
min	70	27	35	88	5	181	24	10	1	5	3	6	
25%	354	3312	6180	347	92	427	717	32	1	30	71	190	
50%	532	3788	10189	466	111	536	1020	42	2	41	184	420	3
75%	802	4594	12952	565	151	943	1261	62	2	60	312	770	6
max	30950	147511	369068	18365	4688	22827	38062	1458	3	2288	8090	25091	152

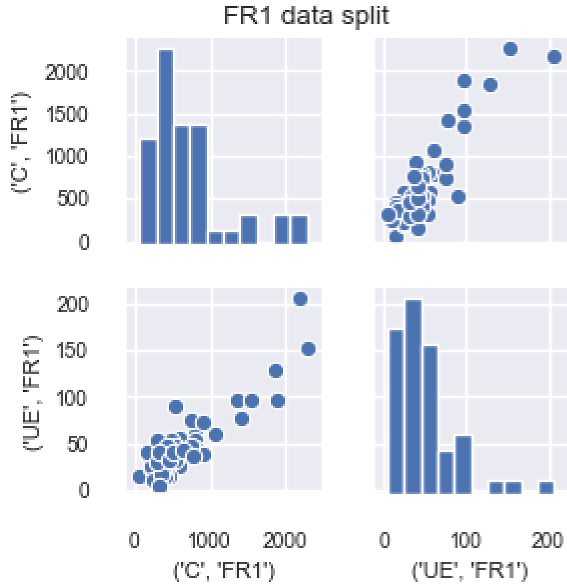
Out[21]:

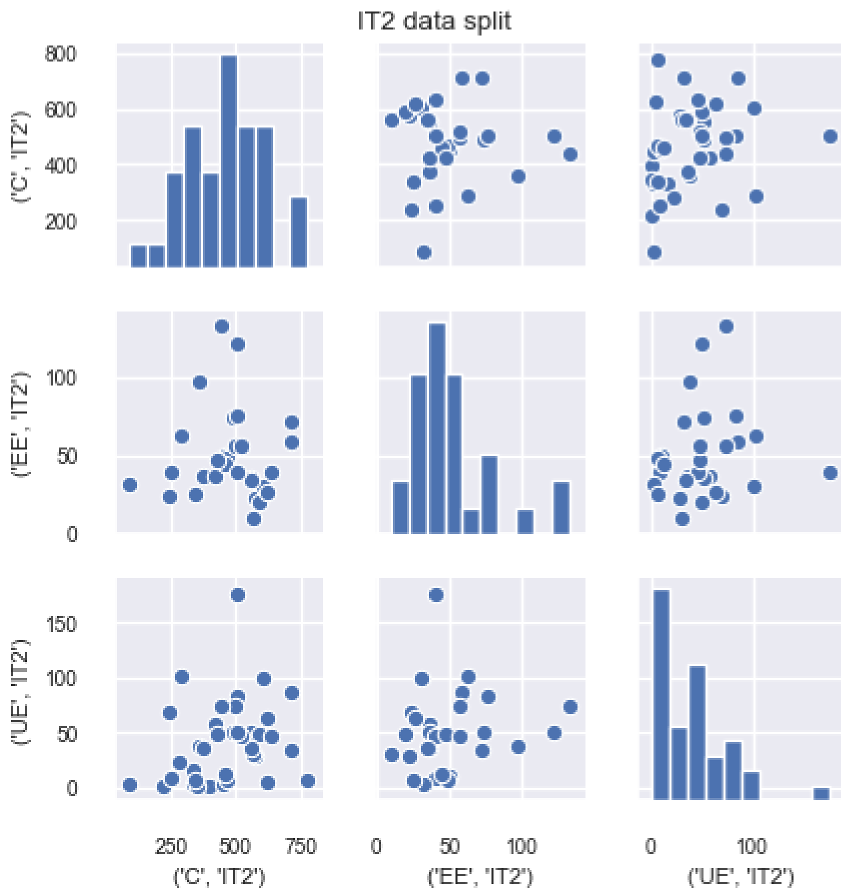
Branch	**FR1**		**GE4**			**HK5**			**IT2**			**SP3**	
Status	**C**	**UE**	**C**	**EE**	**UE**	**C**	**EE**	**UE**	**C**	**EE**	**UE**	**C**	**EE**
count	45	45	42	35	42	42	42	42	41	30	39	42	3
mean	1375	101	7024	1304	385	17574	1812	1194	895	97	78	223	2
std	4540	335	22255	3762	1226	55744	5751	3845	2800	258	240	707	1
min	70	5	27	181	3	35	24	6	88	10	1	5	1
25%	354	30	3312	427	71	6180	717	190	347	32	7	92	1
50%	532	41	3788	536	184	10189	1020	420	466	42	36	111	2
75%	802	60	4594	943	312	12952	1261	770	565	62	60	151	2
max	30950	2288	147511	22827	8090	369068	38062	25091	18365	1458	1527	4688	3

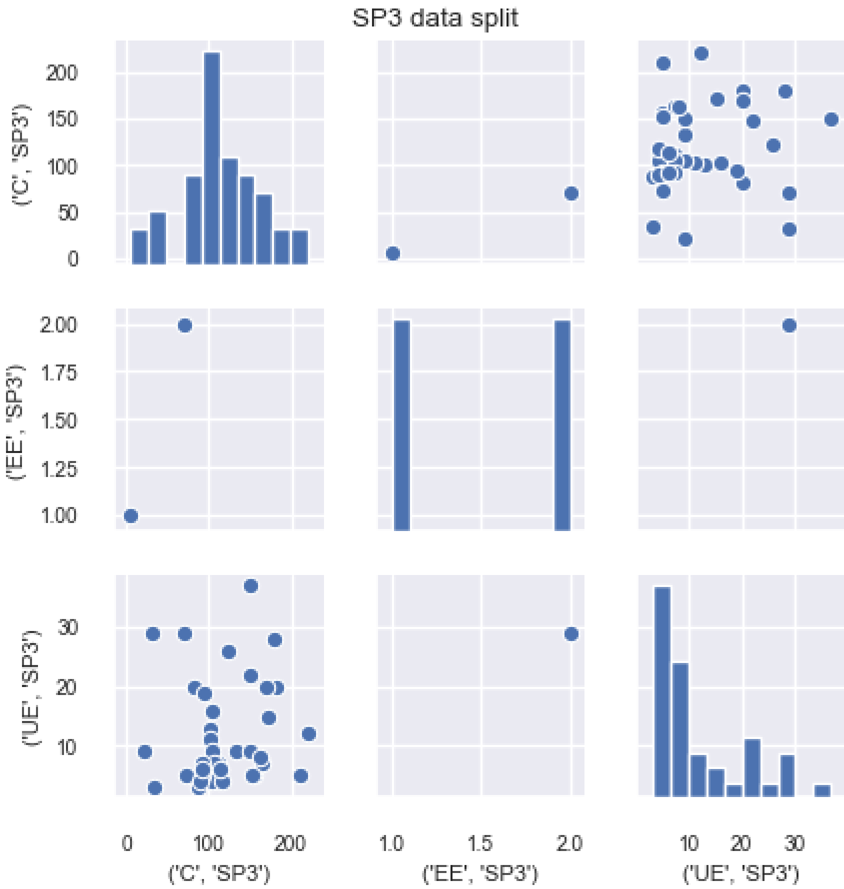
Graphical ditribution and corelation of the data

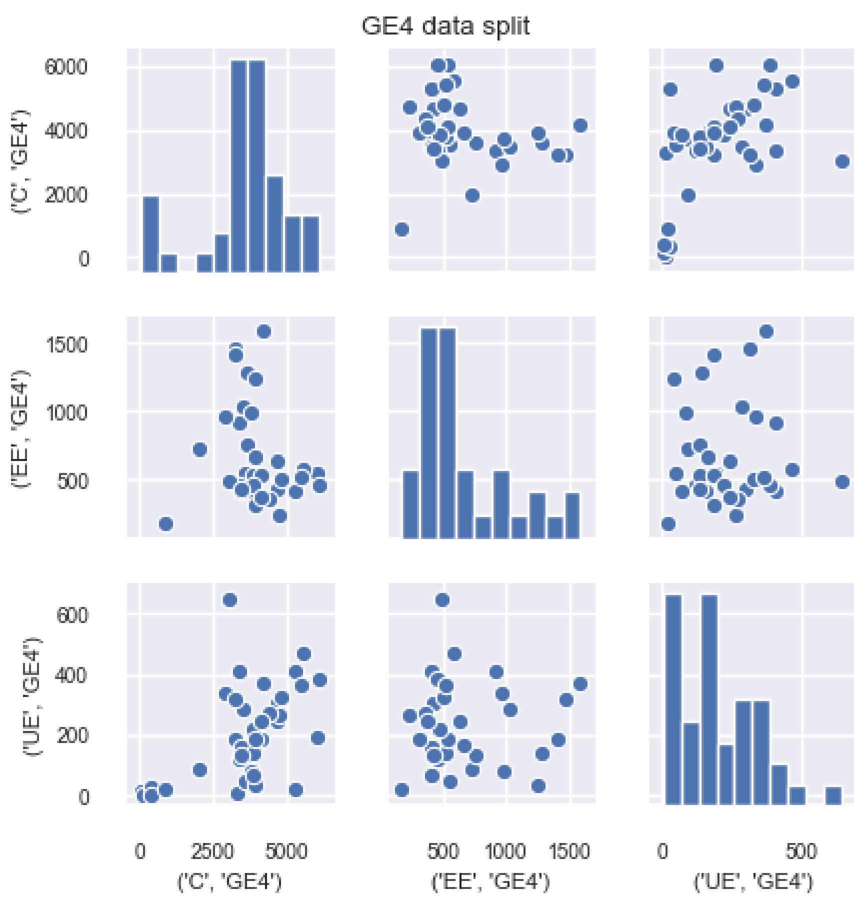
To close on the tasks

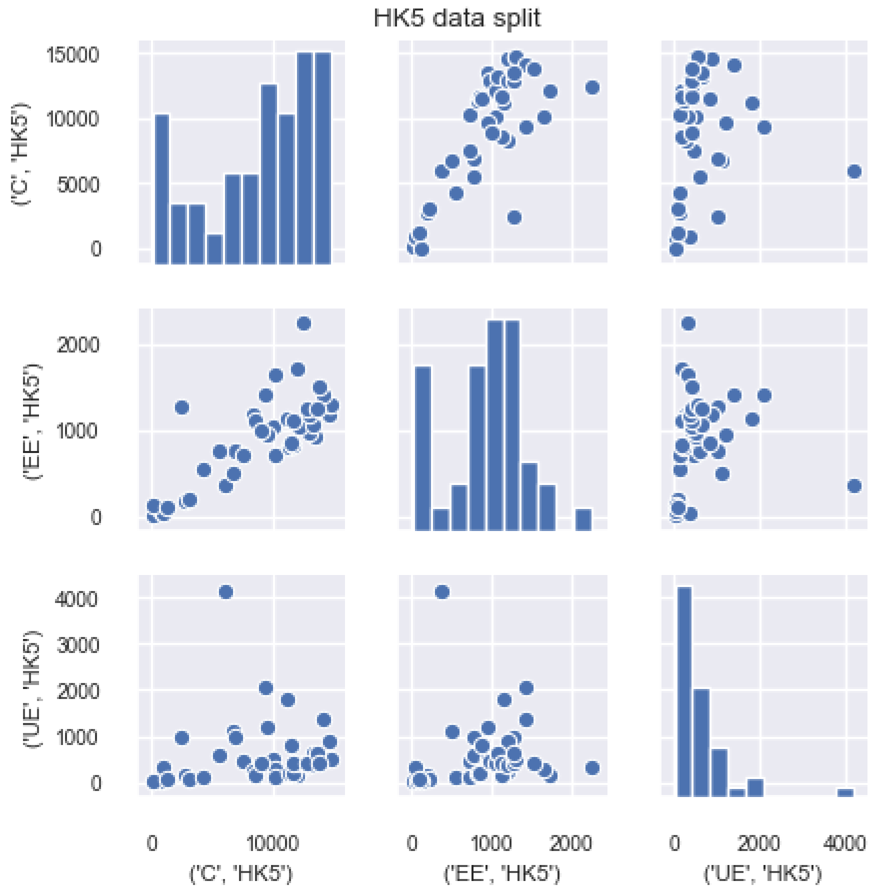
- to determine the type of data and how is its broken down
- identify if the data is normally distributed The charts below help us see how the data is categorized and its correlation.


FR1 data split
('C', 'FR1')
('UE', 'FR1')
0
500
1000
1500
2000
0
50
100
150
200
0
1000
2000
0
100
200


IT2 data split
('C', 'IT2')
('EE', 'IT2')
('UE', 'IT2')
200
400
600
800
0
50
100
0
50
100
150
250
500
750
0
50
100
0
100


SP3 data split
('C', 'SP3')
('EE', 'SP3')
('UE', 'SP3')


GE4 data split
('C', 'GE4')
('EE', 'GE4')
('UE', 'GE4')


HK5 data split
15000
10000
5000
0
2000
1000
0
4000
3000
2000
1000
0
('C', 'HK5')
('EE', 'HK5')
('UE', 'HK5')
0
10000
0
1000
2000
0
2000
4000
('C', 'HK5')
('EE', 'HK5')
('UE', 'HK5')

Forecasting

This is the final task, Using different methods to forecast and draw the results for each one.

- non-seasonal methods (Exponential Smoothing)
  - SES (Simple Exponential Smoothing)
  - Holt's
  - Exponential
  - Additive Damped
  - Multiplicative Damped

- Seasonal methods
  - Additive
  - Multiplicative
  - Additive Damped
  - Multiplicative Damped

As this is a forecast, there is no data to validate against; but, in scenarios where a larger dataset (daily data for two or more years for instance) is available, a subset of the data can be used to forecast (called `Training set` in Machine Learning lingo) and validate the model using the remaining data (`Validation set`) in that way it will be possible to measure the level of accuracy of the different models and determine which one best fits the data provided. (Identify which model has learned better compared to the reallity)

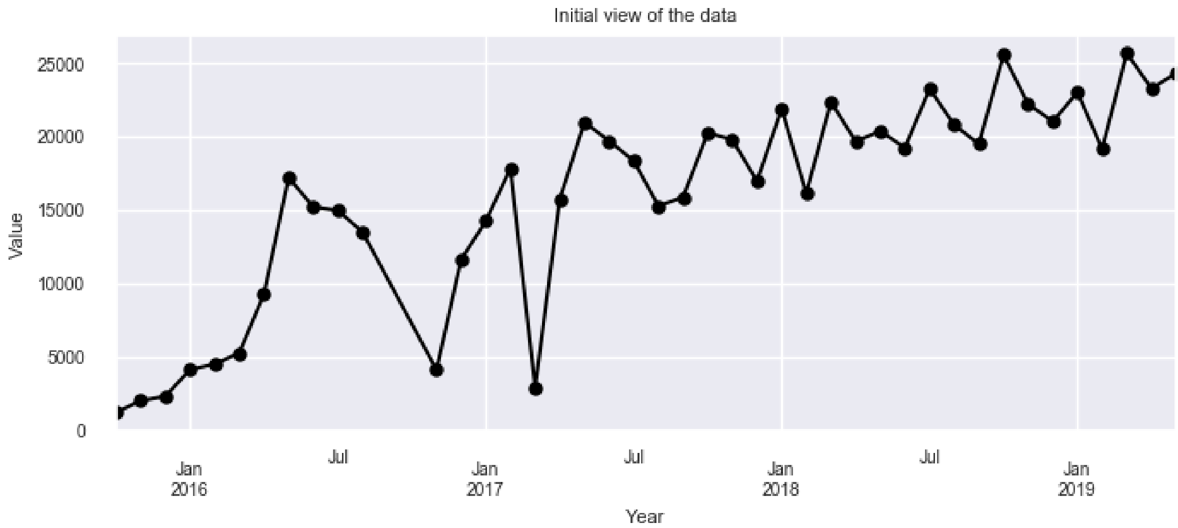
A `Training set` can be 10% of random samples of the data and it's recommended to be no lesser than 50 samples. The information currently availablefor this task is monthly and has only 44 months of information (44 consolidated entries in total) and since the data is quite disperse (as shown in the previous sections) the 10% sample data will be 5 records and usign them to forecast is not the recommended approach.

So, for this particular case Machine Learning will not be used, rather differnet forecast models will be presented and when more data becomes avilable it can be used to run the models again and determine which one predicted the most fitted results.

Exponential smoothing

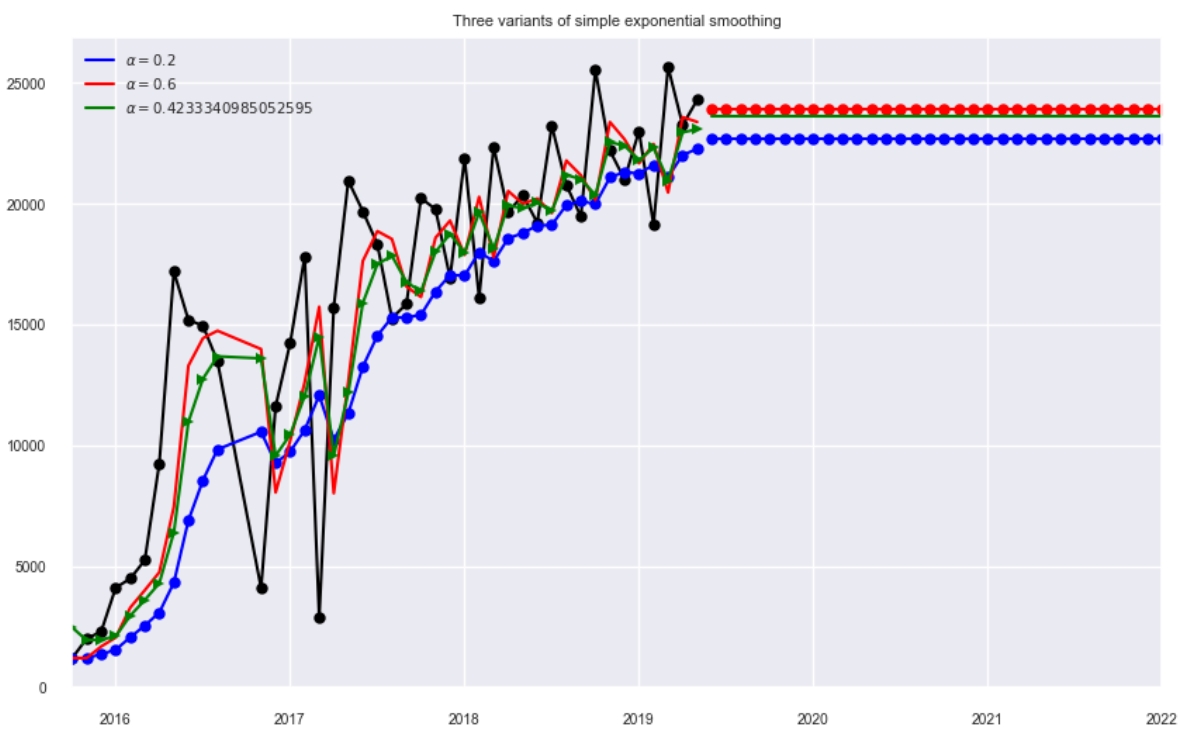
The heavy lifthing was done in the previosus section, at this point we have the data ready for us to apply the different models.

Simple Exponential Smoothing



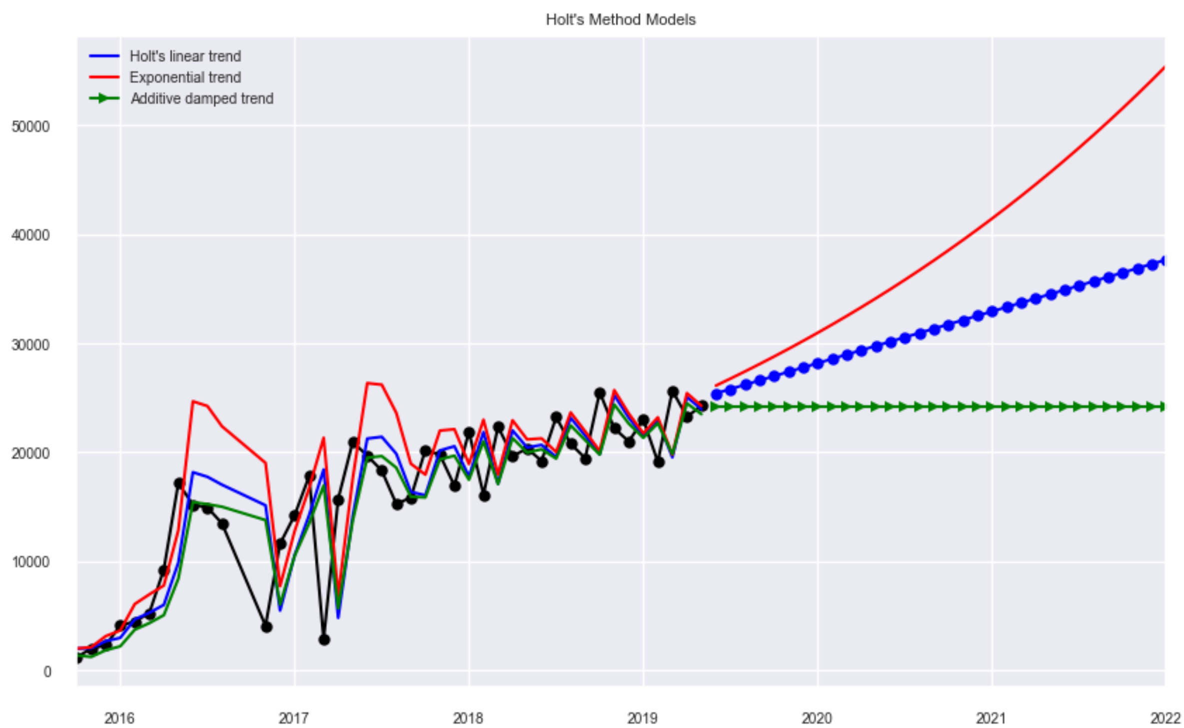
Running three variants of simple exponential smoothing:

1. In `fit1` we do not use the auto optimization but instead choose to explicitly provide the model with the $\alpha = 0.2$ parameter
2. In `fit2` as above we choose an $\alpha = 0.6$
3. In `fit3` we allow statsmodels to automatically find an optimized α value for us. This is the recommended approach.



Fitting using Holt's Method

1. In `fit1` we again choose not to use the optimizer and provide explicit values for $\alpha = 0.8$ and $\beta = 0.2$
2. In `fit2` we do the same as in `fit1` but choose to use an exponential model rather than a Holt's additive model.
3. In `fit3` we used a damped versions of the Holt's additive model but allow the dampening parameter ϕ to be optimized while fixing the values for $\alpha = 0.8$ and $\beta = 0.2$



Seasonally adjusted data

Fitting five Holt's models.
The below table allows us to compare results when we use exponential versus additive and damped versus non-damped.

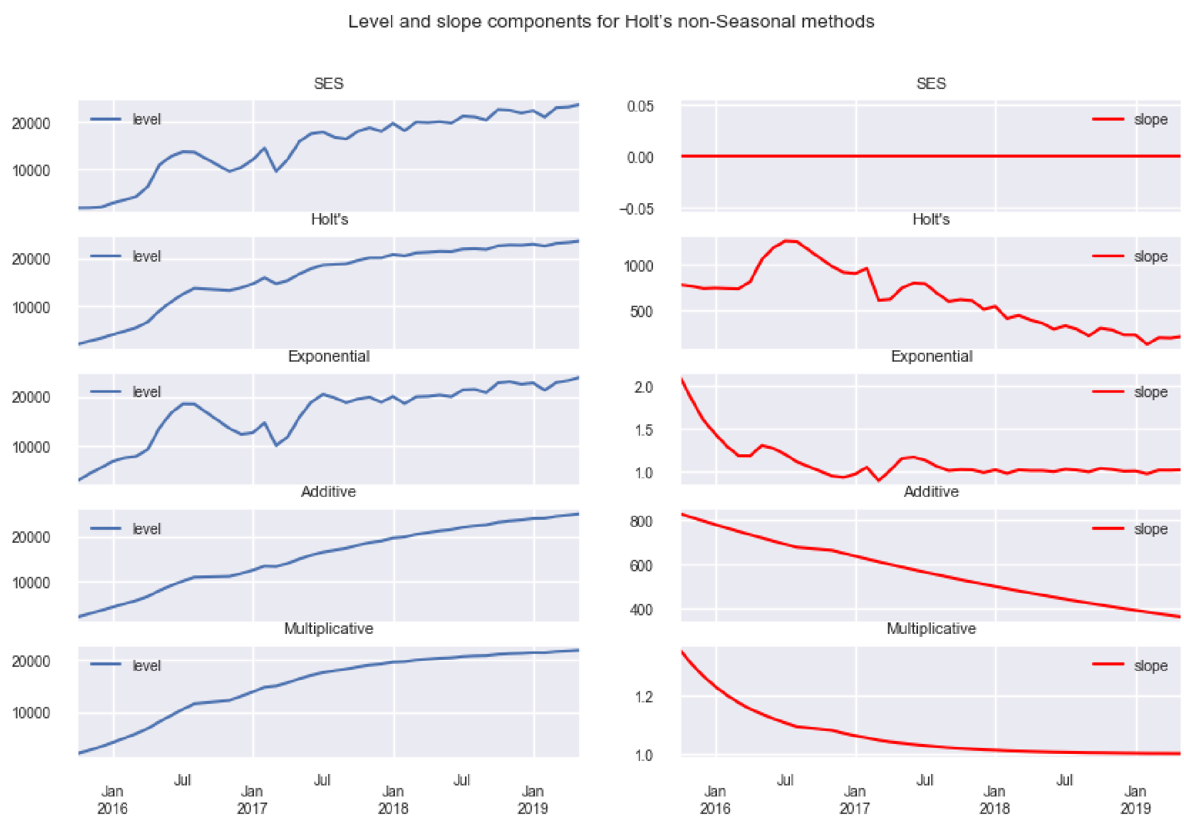
Note: `fit4` does not allow the parameter ϕ to be optimized by providing a fixed value of $\phi = 0.98$

Parameters generated by the non-seasonal models

Out[28]:

	SES	Holt's	Exponential	Additive	Multiplicative
α	0.42	0.16	0.43	0.06	0.04
β	nan	0.16	0.43	0.00	0.04
ϕ	nan	nan	nan	0.98	0.88
l_0	2,469.37	1,196.00	1,739.48	1,201.99	1,595.84
b_0	nan	804.00	2.41	828.90	1.35
SSE	736,359,506.17	645,065,499.63	1,145,410,990.34	577,153,213.82	550,409,078.07

The following plots can be used to evaluate the level and slope/trend components of the above table's fits.

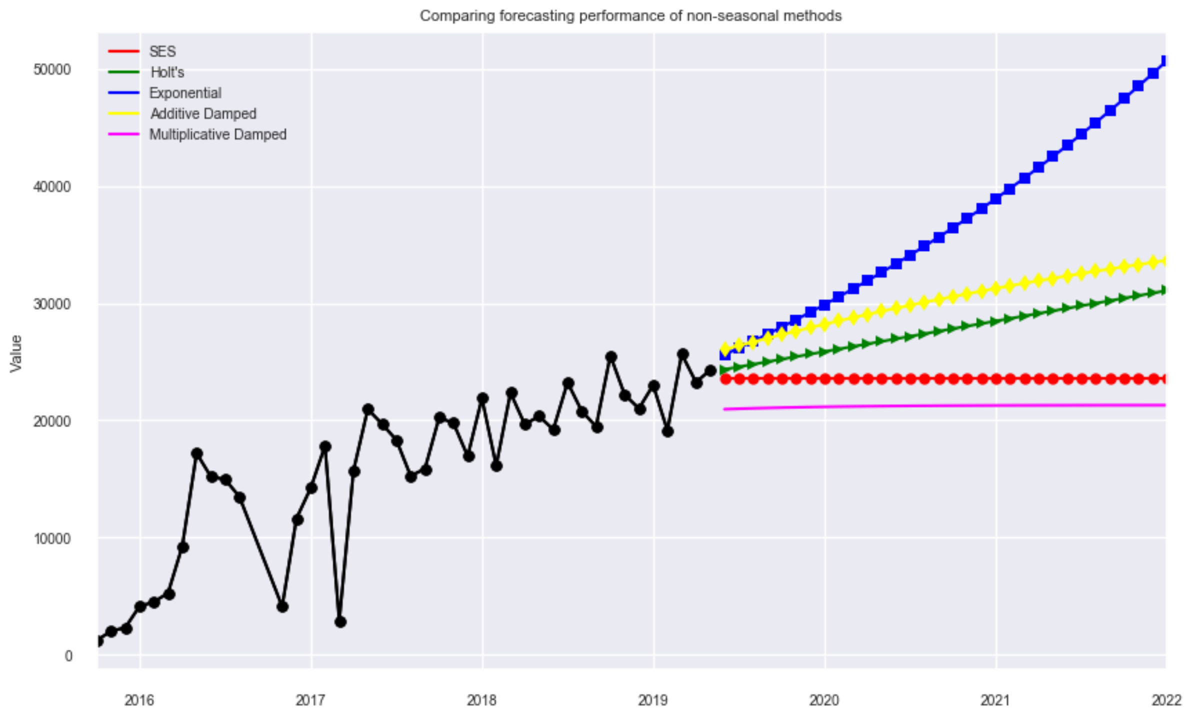



```
Level and slope components for Holt’s methods
```

note: The seasonal component for the above forecast is zero.

Comparison

Comparing the Simple Exponential Smoothing and Holt's Methods for various additive, exponential and damped combinations. All of the models parameters will be optimized by statsmodels.




```
Comparing forecasting performance of non-seasonal methods.
```

Plots of Seasonally Adjusted Data

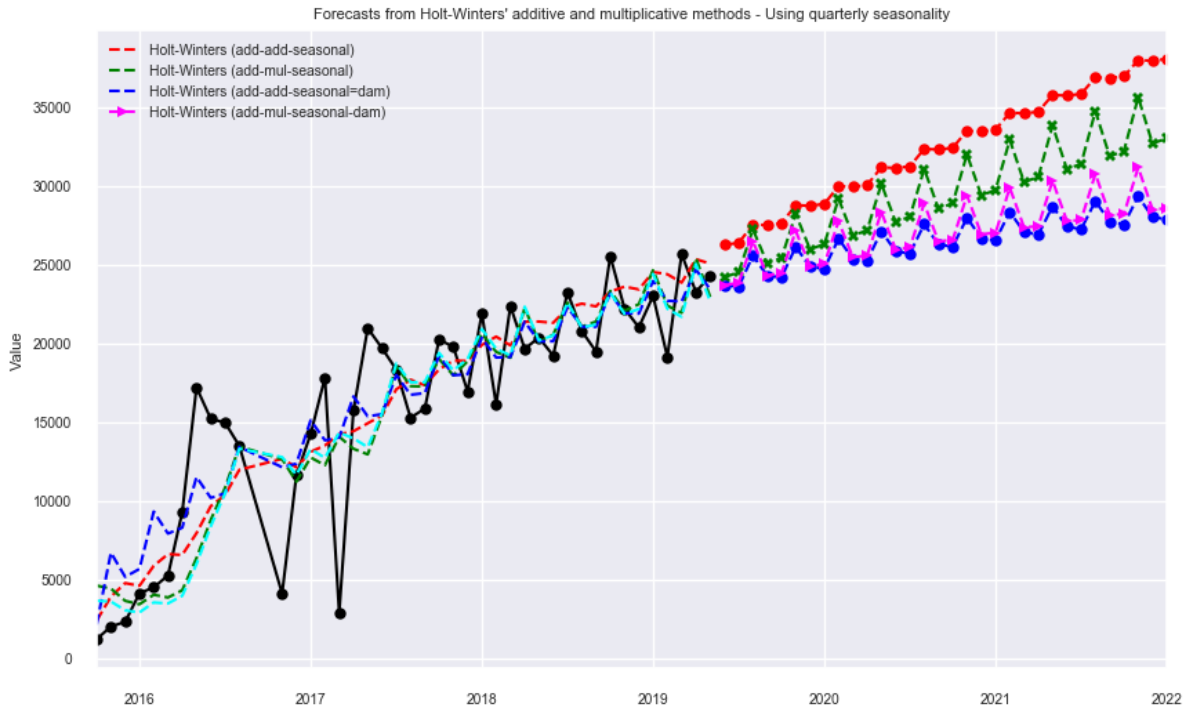
Holt's Winters Seasonal

Finally we are able to run full Holt's Winters Seasonal Exponential Smoothing including a trend component and a seasonal component. statsmodels allows for all the combinations including as shown in the examples below:

1. `fit1` additive trend, additive seasonal of period `season_length=4` and the use of a Box-Cox transformation.
2. `fit2` additive trend, multiplicative seasonal of period `season_length=4` and the use of a Box-Cox transformation..
3. `fit3` additive damped trend, additive seasonal of period `season_length=4` and the use of a Box-Cox transformation.
4. `fit4` additive damped trend, multiplicative seasonal of period `season_length=4` and the use of a Box-Cox transformation.

The plot shows the results and forecast for `fit1` to `fit4` . The table allows us to compare the results and parameterizations.

```
C:\ProgramData\Anaconda3\lib\site-packages\statsmodels\tsa\holtwinters.py:71
1: ConvergenceWarning: Optimization failed to converge. Check mle_retvals.
  ConvergenceWarning)
```

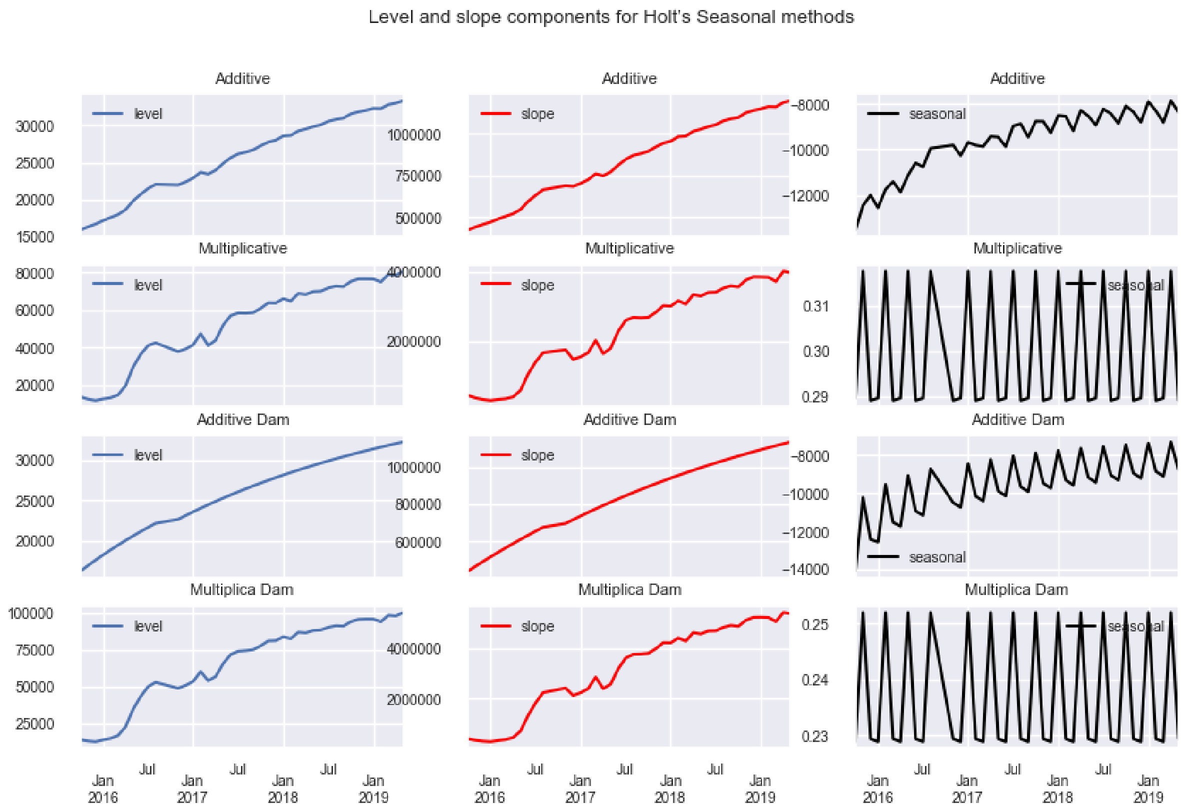



```
Forecasting using Holt-Winters method with both additive and multiplicative s
easonality.
```

Parameters generated by the seasonal models

Out[32]:

	Additive	Multiplicative	Additive Dam	Multiplica Dam
α	0.09	0.24	0.00	0.21
β	0.00	0.06	0.00	0.12
ϕ	nan	nan	0.98	0.98
γ	0.03	0.00	0.00	0.00
l_0	422,135.75	421,731.25	426,992.13	421,731.25
b_0	19,396.05	19,245.80	27,548.84	18,895.68
SSE	558,982,601.04	634,024,207.62	467,562,310.69	637,054,498.24




```
Level and slope components for Holt’s methods
```

Looking at the levels, slopes/trends and seasonal components of the models.

The Internals

Following is a sample of the tables generated during the modeling that show side by side the original values y_t, the level l_t, the trend b_t, the season s_t and the fitted values $\hat{y}_t$.

Out[35]:

	$\hat{y}_t$	b_t	l_t	s_t	y_t
2015-10	2370.732598	425679.310740	15852.484143	-13528.257294	1196.0
2015-11	3898.464586	442883.418399	16267.445598	-12458.910816	2000.0
2015-12	4759.131859	458431.471828	16643.939355	-12008.417359	2296.0
2016-01	4585.124155	472542.551995	17114.266105	-12554.714548	4112.0
2016-02	5853.214548	490520.459816	17526.955153	-11751.874660	4485.0
2016-03	6601.586963	506409.224557	17932.785963	-11411.889057	5248.0
...	...	...	...	...	...
2021-08	36888.601625	NaN	NaN	NaN	NaN
2021-09	36851.629768	NaN	NaN	NaN	NaN
2021-10	36954.104015	NaN	NaN	NaN	NaN
2021-11	37990.969122	NaN	NaN	NaN	NaN
2021-12	37954.405314	NaN	NaN	NaN	NaN
2022-01	38055.749927	NaN	NaN	NaN	NaN

74 rows × 5 columns

Out[36]:

	$\hat{y}_t$	b_t	l_t	s_t	y_t
2015-10	4599.409156	427472.812048	13504.239496	0.289554	1196.0
2015-11	4412.820573	354417.108209	12265.270675	0.317649	2000.0
2015-12	3633.051242	307968.468222	11538.026060	0.289034	2296.0
2016-01	3413.229474	280098.170059	12376.632754	0.289554	4112.0
2016-02	4019.624683	309465.069015	13007.735646	0.317649	4485.0
2016-03	3844.426420	330866.868287	14507.798763	0.289034	5248.0
...	...	...	...	...	...
2021-08	34738.917087	NaN	NaN	NaN	NaN
2021-09	31881.428172	NaN	NaN	NaN	NaN
2021-10	32210.405836	NaN	NaN	NaN	NaN
2021-11	35632.720981	NaN	NaN	NaN	NaN
2021-12	32692.138937	NaN	NaN	NaN	NaN
2022-01	33020.034281	NaN	NaN	NaN	NaN

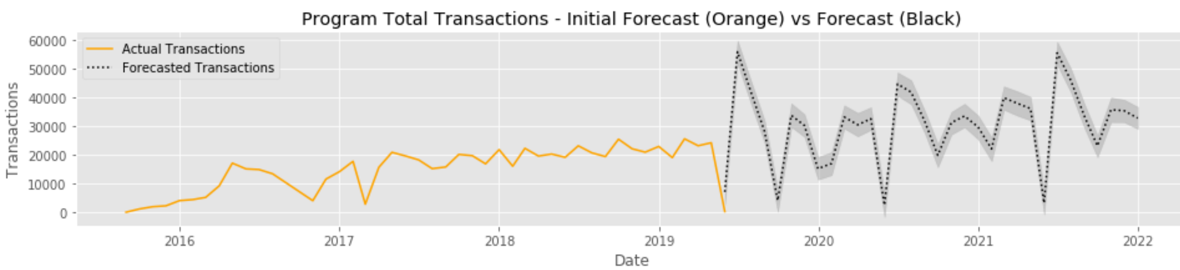
74 rows × 5 columns

Others...

There are other methods that can be used to forecast, for the purpose of this activity below are some examples please note: **`(the charts below are not automatically updated when refreshing this book. these are here solely to give visibility of other options and a taste of the results)`**

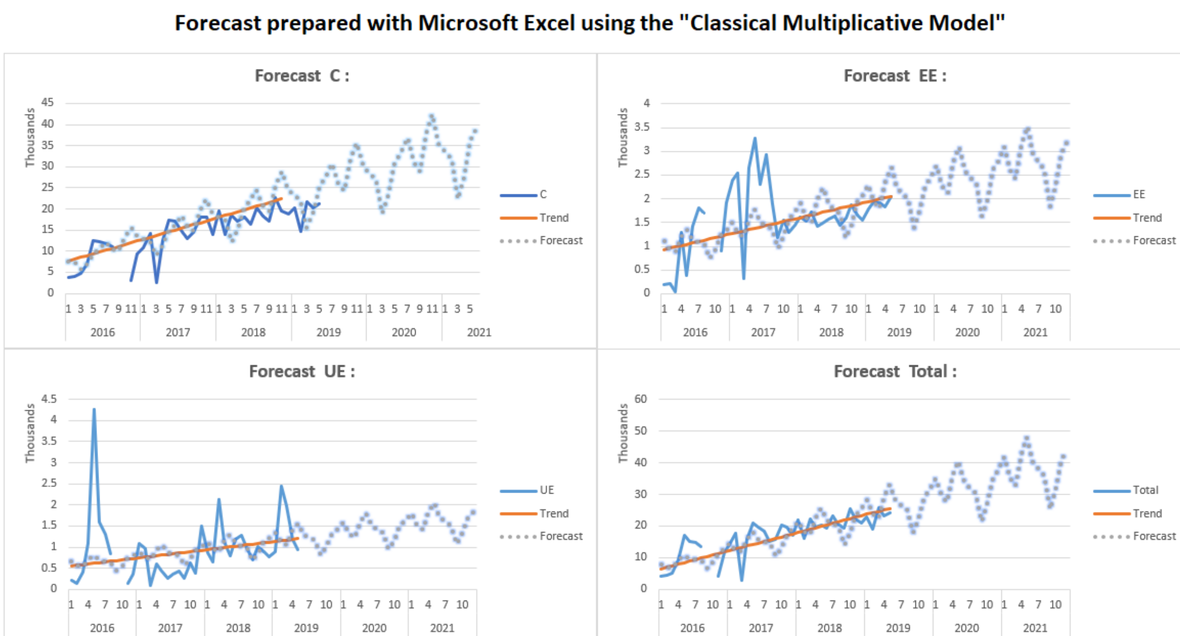
using Facebook Prophet forecasting package as the basis and customizing the final chart.

Out[37]:



Forecast prepared with Microsoft Excel using the "Classical Multiplicative Model"

Out[38]:



This is the end of the forecast and the activity