416-580-4307
kpk.cbe@gmail.com
Karthik Palaniswamy, B.Tech, EIT, PMP
Data Analyst
https://www.linkedin.com/in/kpal/
https://github.com/kpkpal

OUTLIERS AND ANOMALY DETECTION

USING

POWERBI DAX

AND

PYTHON MACHINE LEARNING

Table of Contents

416-580-4307
kpk.cbe@gmail.com
Karthik Palaniswamy, B.Tech, EIT, PMP
Data Analyst
https://www.linkedin.com/in/kpal/
https://github.com/kpkpal

1. MACHINE LEARNING ANOMALY DETECTION IN POWERBI

This project detects anomaly using Isolation Forest method using Machine Learning with PowerBI DAX formulas and Python scrips to analyze the temperature parameter of a machine.

The main idea, which is different from other popular outlier detection methods, is that Isolation Forest explicitly identifies anomalies instead of profiling normal data points. Isolation Forest, like any tree ensemble method, is built on the basis of decision trees. In these trees, partitions are created by first randomly selecting a feature and then selecting a random split value between the minimum and maximum value of the selected feature.

The Isolation Forest 'isolates' observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature.

Backgroud of the Dataset used:

The dataset contains day wise temperature recording for a machine. The anomaly detection is performed for the temperature column in the dataset. A sample of the data used is provided below.

Sample source database table: Source: Github, Type: CSV, Columns: 2, Rows: 1769

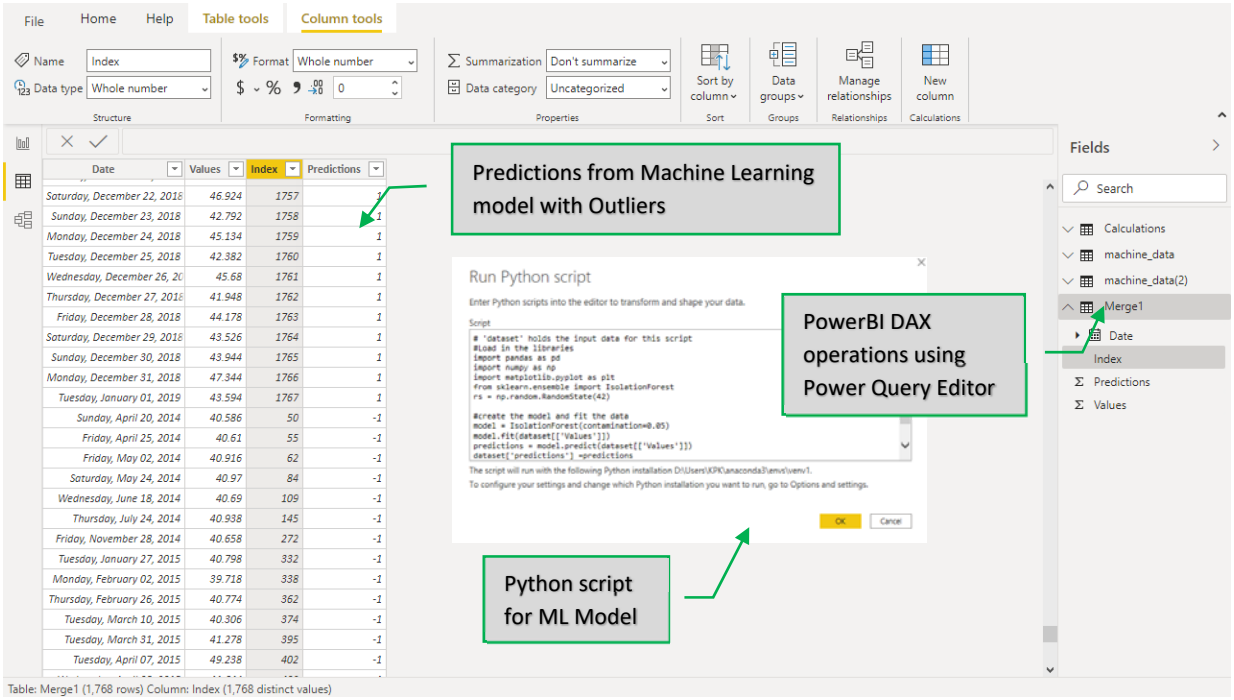
Date	Temperature
03/01/2014	45.868
03/02/2014	47.606
03/03/2014	42.58
03/04/2014	46.03
03/05/2014	44.992
03/06/2014	45.238
03/07/2014	45.752
03/08/2014	46.476
03/09/2014	42.752
03/10/2014	46.156
03/11/2014	43.35
03/12/2014	46.292
03/13/2014	43.428
03/14/2014	46.314
03/15/2014	43.688
03/16/2014	45.4

The above raw data was extracted and transformed in PowerBI for further analysis using Python code. The SKlearn Machine Learning module was used to determine the outliers using the Isolation Forest method. The outliers identified are represented as (-1) under the Predictions column in the below table.

416-580-4307
kpk.cbe@gmail.com

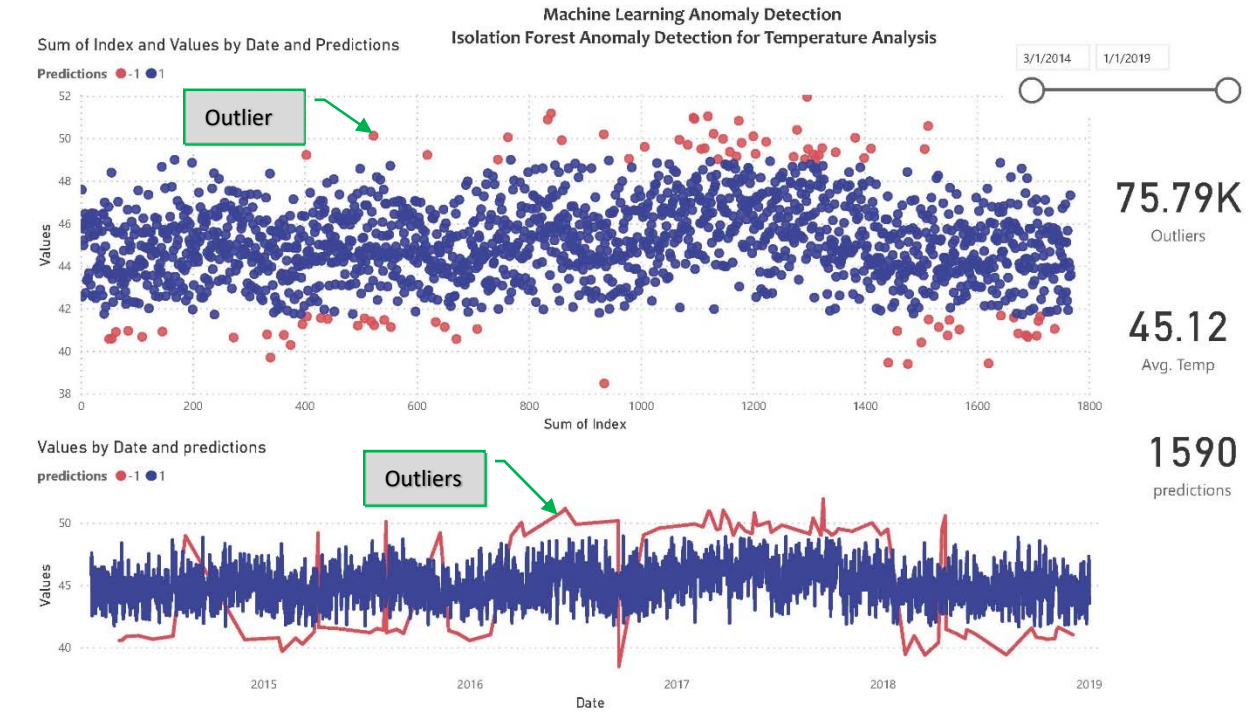
Karthik Palaniswamy, B.Tech, EIT, PMP
Data Analyst

https://www.linkedin.com/in/kpal/
https://github.com/kpkpal



The predicted output values are plotted in PowerBI to visually represent the outliers in the dataset. The below dashboard provides a dynamic visualization of the complete dataset.

MACHINE LEARNING ANOMALY DETECTION



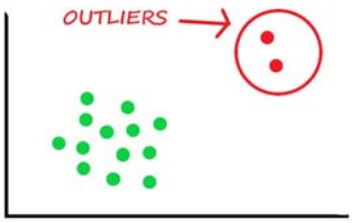
416-580-4307
kpk.cbe@gmail.com
Karthik Palaniswamy, B.Tech, EIT, PMP
Data Analyst
https://www.linkedin.com/in/kpal/
https://github.com/kpkpal

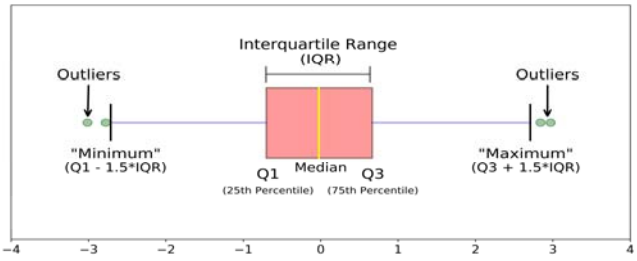
2. STATISTICAL OUTLIER DETECTION IN TIME SERIES DATA

Outlier detection or Anomaly detection is the identification of rare items, events or observations which rise suspicions by differing significantly from the majority of the data. Outlier detection is the process of detecting and excluding them from a given set of data. Typical anolomous problems in a real world could be:

- A bank fraud
- Medical problems
- A structural defect
- Errors in a text, measurement, etc

Outliers can be explained as the data point that is more than 1.5 times the Inter Quartile Range (IQR) where Q1 is the median of the lower half of the data, and Q3 is the median of the upper half of the data.





Microsoft PowerBI provides anomaly detection and statistical analysis methods that can be applied to a dataset. The result of the analysis can be translated as a report and a dashboard visual for easy detection. The analysis also provides a deeper insight into the business. A typical data was used to identify the outliers in a dataset using PowerBI and the results are presented below.

Backgroud of the Dataset used:

The dataset contains day wise users index for a newly launched website. The number of users visiting the website are plotted against the day index. A sample of the data used is provided below.

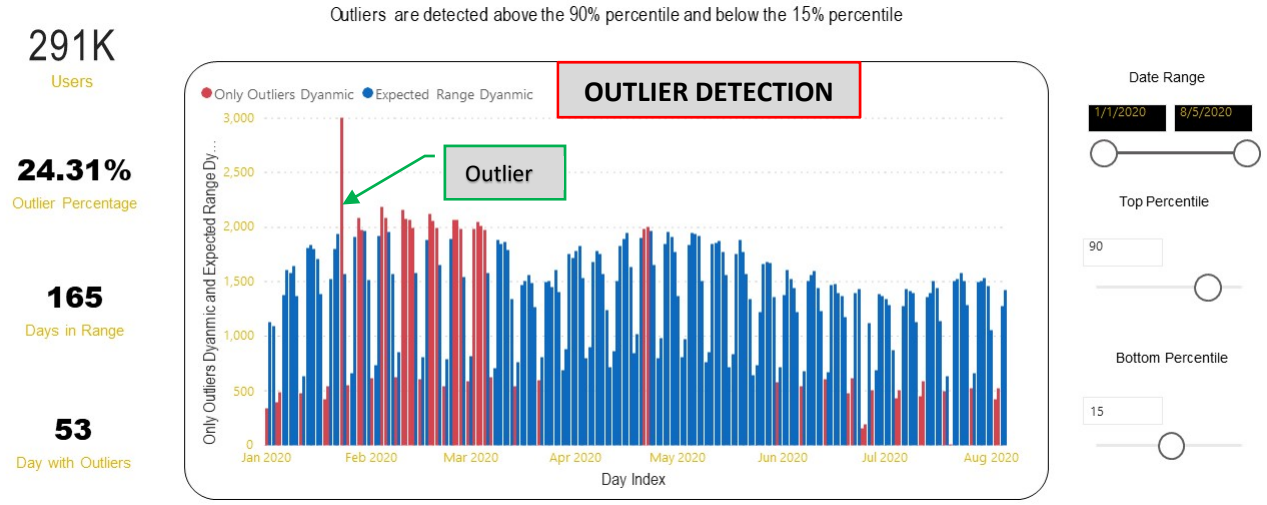
Sample database table: Source: Github, Type: CSV, Columns: 1, Rows: 222

DayIndex	Users
01/01/2020 0:00	335
01/02/2020 0:00	1131
01/03/2020 0:00	1094
01/04/2020 0:00	393
01/05/2020 0:00	487
01/06/2020 0:00	1380
01/07/2020 0:00	1607
01/08/2020 0:00	1578
01/09/2020 0:00	1640

416-580-4307
kpk.cbe@gmail.com

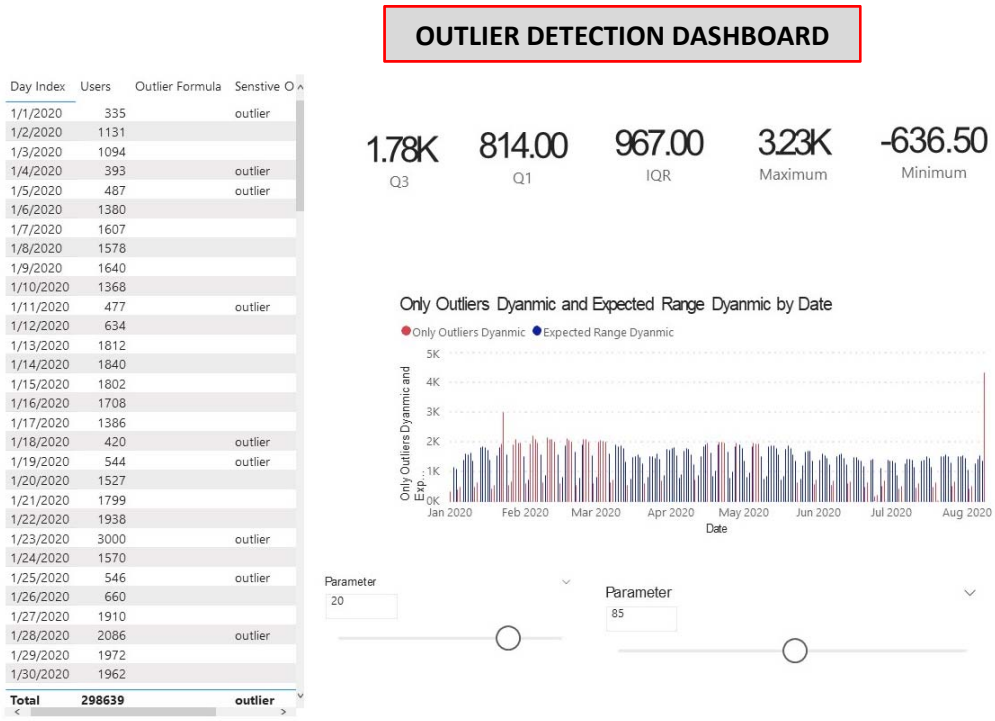
Karthik Palaniswamy, B.Tech, EIT, PMP
Data Analyst

https://www.linkedin.com/in/kpal/
https://github.com/kpkpal



In the above chart, the outliers are calculated in PowerBI from the given dataset and they are represented as red coloured bars.

Dynamic dashboard with table view developed in PowerBI.



Day Index	Users	Outlier Formula	Senstive O
1/1/2020	335		outlier
1/2/2020	1131		
1/3/2020	1094		
1/4/2020	393		outlier
1/5/2020	487		outlier
1/6/2020	1380		
1/7/2020	1607		
1/8/2020	1578		
1/9/2020	1640		
1/10/2020	1368		
1/11/2020	477		outlier
1/12/2020	634		
1/13/2020	1812		
1/14/2020	1840		
1/15/2020	1802		
1/16/2020	1708		
1/17/2020	1386		
1/18/2020	420		outlier
1/19/2020	544		outlier
1/20/2020	1527		
1/21/2020	1799		
1/22/2020	1938		
1/23/2020	3000		outlier
1/24/2020	1570		
1/25/2020	546		outlier
1/26/2020	660		
1/27/2020	1910		
1/28/2020	2086		outlier
1/29/2020	1972		
1/30/2020	1962		
Total	**298639**		**outlier**

Conclusion:

Outlier information is very useful when data is compared with the original data. The above identification of outlliers from the original dataset provides an opportunity to further investigate the causation using PowerBI.

416-580-4307
kpk.cbe@gmail.com
Karthik Palaniswamy, B.Tech, EIT, PMP
Data Analyst
https://www.linkedin.com/in/kpal/
https://github.com/kpkpal

3. ANOMALY DETECTION IN SUPERSTORE SALE DATA

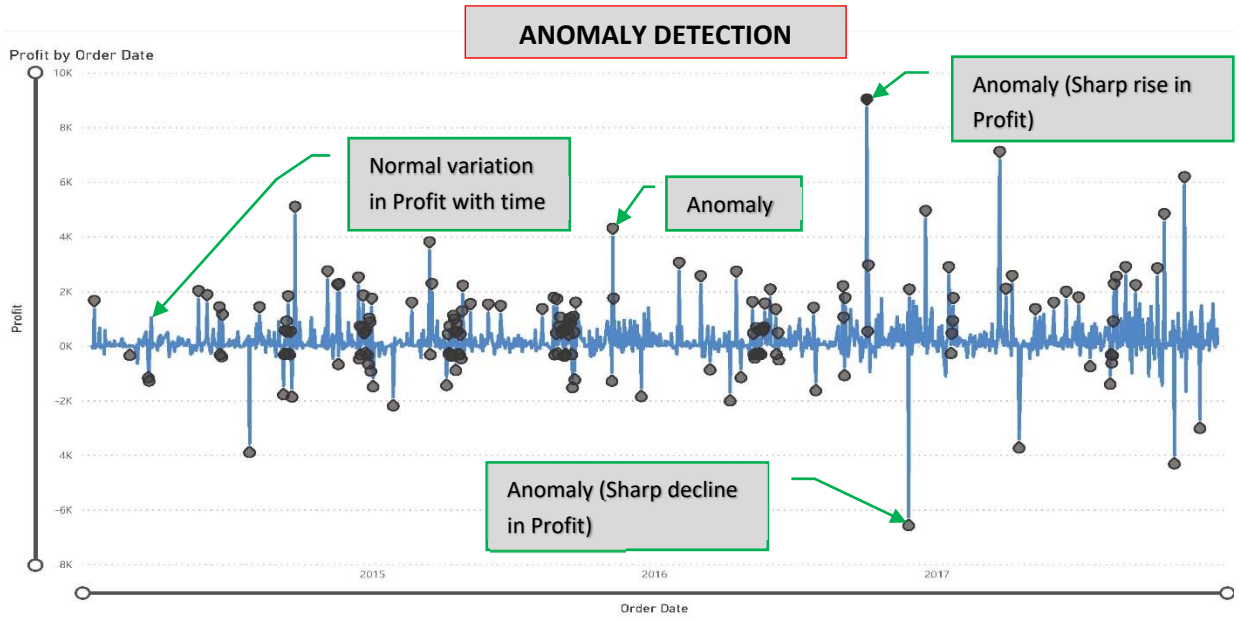
Anomaly detection (or outlier analysis) is referred as something that deviates from what is standard, normal, or expected. Anomalous data can indicate critical incidents, such as a fraudlent transaction, technical glitch, or potential opportunities, or a change in consumer behavior. Identifying outliers and correlating with various influencing factors can deliver insights to business decision makers.

Backgroud of the Dataset used:

The dataset contains the sale of consumer products by a superstore in USA. The Order date and Profits of the superstore was analysed to understand the anomalies in the rise and fall of profits with respect to time of goods purchased. Two sharp in the data points, one for sharp increase in profit one for sharp decline in profit were selected for further analysis.

Sample database table: Source: Github, Type: Excel, Columns: 21, Rows: 9995, Columns analysed: Order date and Profit

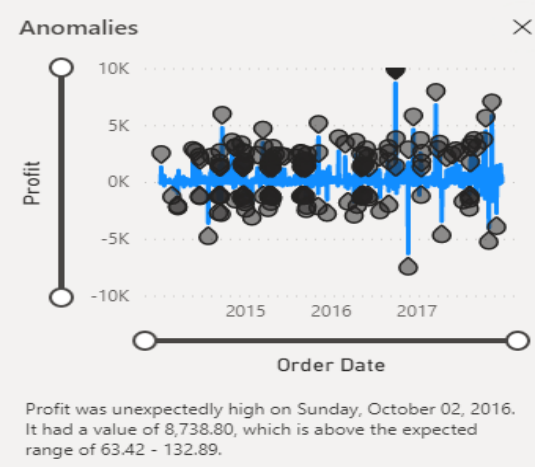
Row ID	Order Date	Ship Date	Postal Code	Region	Category	Sales	Quantity	Discount	Profit
43	Sunday, July 17, 2016	Friday, July 22, 2016	90049	West	Office Supplies	77.88	2	0	3.894
514	Thursday, December 21, 2017	Monday, December 25, 2017	90049	West	Office Supplies	6.63	3	0	1.7901
....	....	....	....	....	....	....	....	....	....

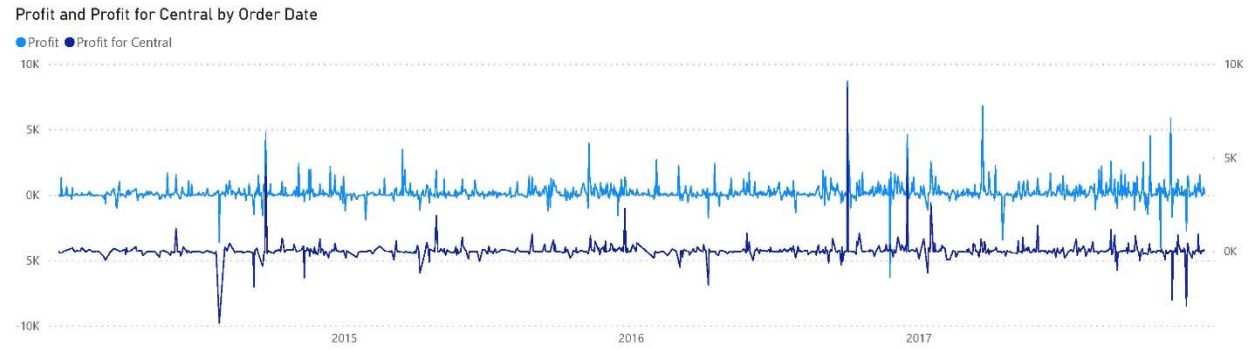


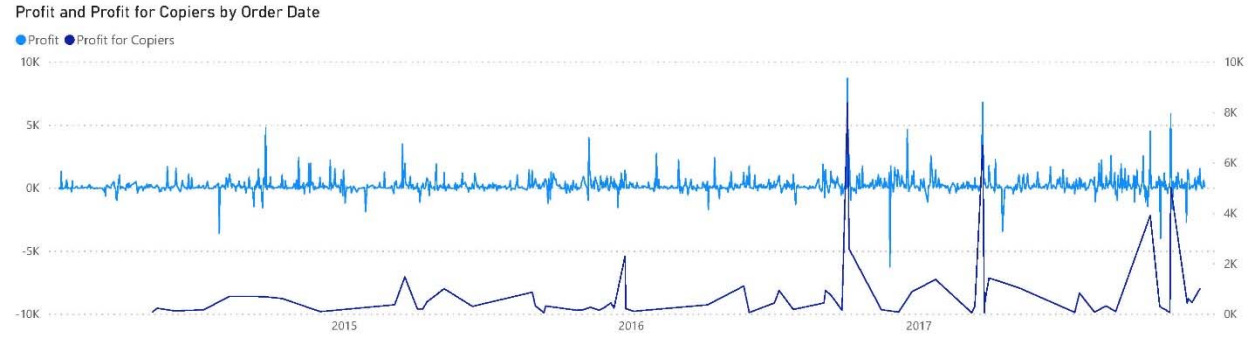
416-580-4307
kpk.cbe@gmail.com
Karthik Palaniswamy, B.Tech, EIT, PMP
Data Analyst
https://www.linkedin.com/in/kpal/
https://github.com/kpkpal

Rise in Profit – A Deeper Analysis using PowerBI

When the data point was selected, the reasons contributing to the anomaly were available for further investigation. The sharp rise in Profit for the identified above was found to be contributed by two major reasons (Central region and Copiers product sub-category). Their contribution is more than 50% for the rise in profit. The detailed graphs and the findings are presented below.







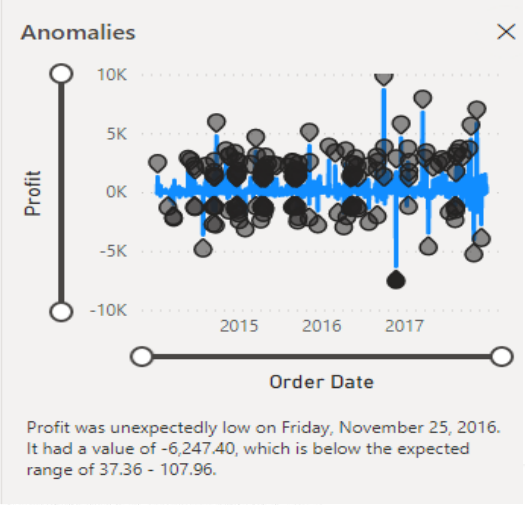
416-580-4307
kpk.cbe@gmail.com

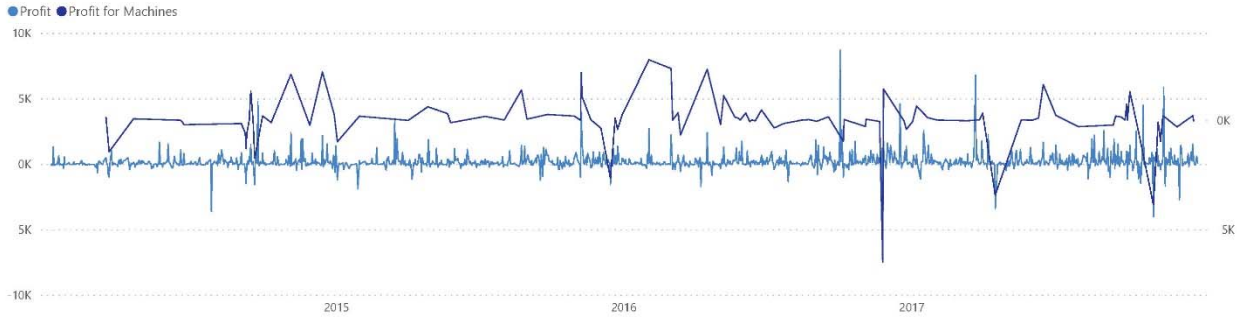
Karthik Palaniswamy, B.Tech, EIT, PMP
Data Analyst

https://www.linkedin.com/in/kpal/
https://github.com/kpkpal

Profit decline by Sub-category and State-wise Comparison:

Similar to the above the reasons two factors were found as the reasons for sharp decline in Profit in 2016. These two factors, Ohio state and product Machines under the Sub-Category were responsible for more than 50% of the total Profit decline. The detailed graphs and the findings are presented below.





Conclusion:

From the above charts and analysis, the underlying reasons and their significant impact to the business can be well correlated. The knowledge gained from this analysis could be further leveraged to promote the business interest.