GrowHaus Report

Moran Wang & Jingru Ma

March 12, 2018

Objective:

Our project is a database about The GrowHaus. It is a non-prot urban green house who focus on food production, distribution, and education.The company is located in northern Denver and provide lots of their services for Denver. The project shows data of three years, 2015,2016 and 2017.

By analyzing the data of these three years, the company can understand their production better and predict revenue based on some predictors. To help key stakeholders, we have created revenue KPI as references. We also plan to to implement A an SQL based database and a web dash for real time analysis of these metrics and SWOT analysis.

Data:

- The name of the dataset is 2017 Update Weekly Metrix Compilation
- There are 12 sheets in the excel file
- We focus on Service Learning, Food boxes and Market
- Service learning is the company's food education branch, including information about students' education and revenues
- Food box includes the information about the type of food box and number of food boxes that was delivered to three different zones. It is in company's distribution branch.
- Market includes the information about sales and revenue from each zone. It is also in company's distribution branch.

Tools, libraries and language:

- R
- library: tidyverse, realxl, dplyr, zoo, ggplot2,
- Code by myself or others: import_tidy_function, BulletGraph.R

Pre-processing:

- We save the xlsx file to csv in order to get easy well controlled date information.

Evaluation:

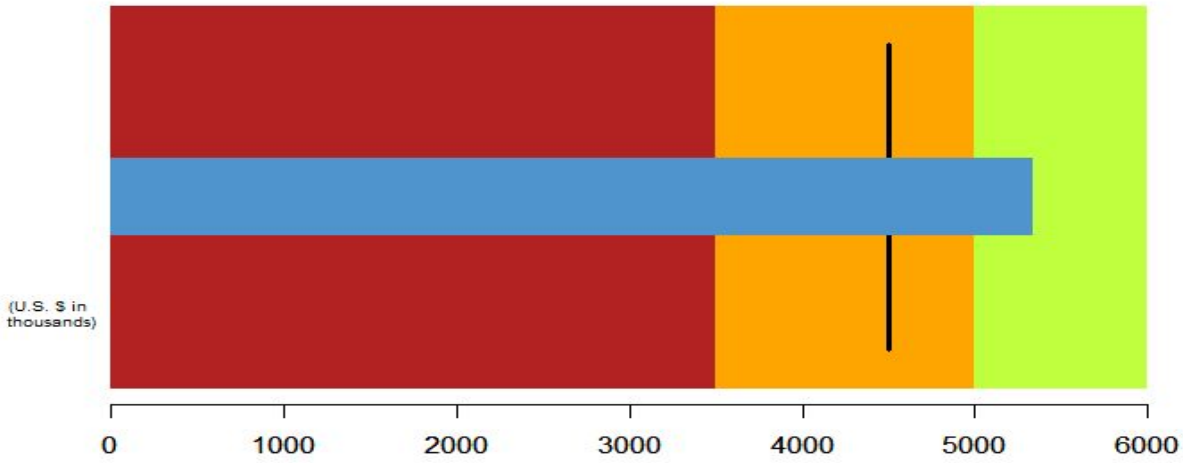
- Data Cleaning activity (in R):
    - Converting date from factor to date format in order to manipulate the date and draw time series graph

    - Removing all $ signs from original data.
    - Removing all redundant string information
- About missing values:
    - Checking missing values
    - Deleting missing values when each row is missing

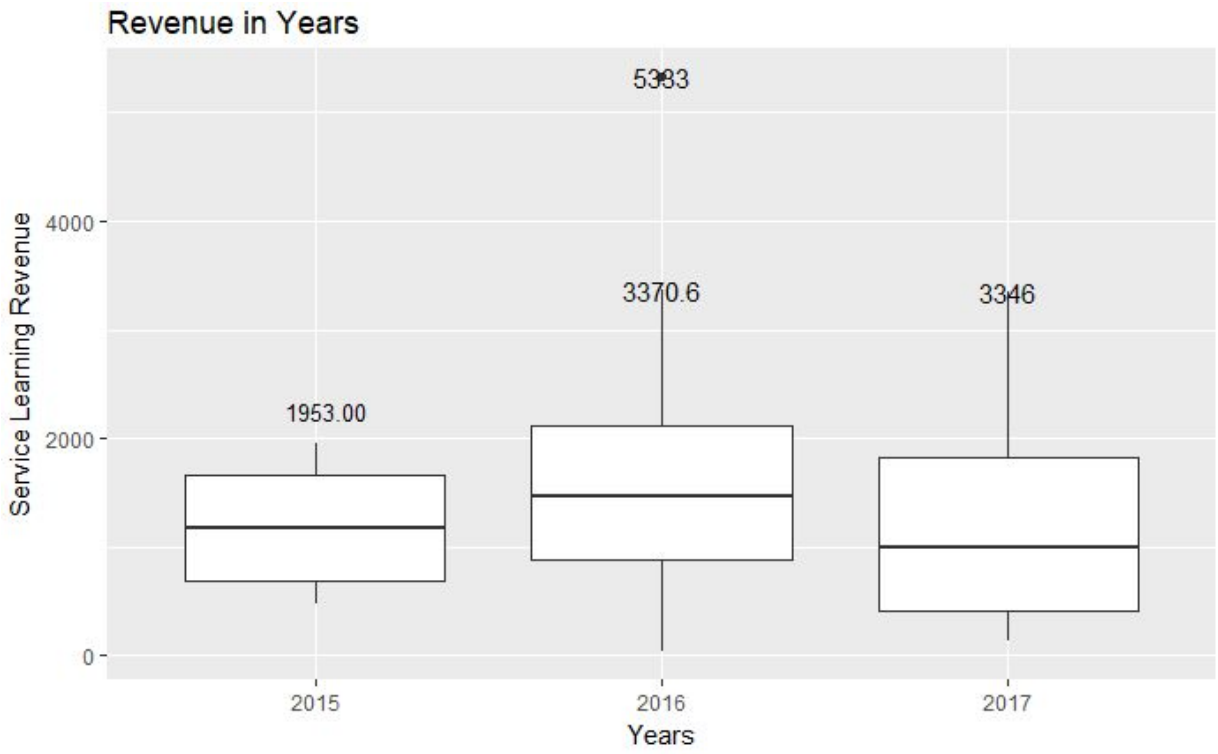
Service Learning

Data Visualization/summary:

Service Learning Revenue KPI



The KPI of service learning shows the revenue of service learning graphically. This graph is used to evaluate the level of revenue. The reference point is 4500 and the three levels of background are 3500,5000 and 6000. These three level in background represent bad, good and excellent level. The blue bar reach to the green level which is the excellent level. The critical black line is the target value. The target value is 4500.



This graph is showing the service learning revenue in the year of 2015,2016 and 2017 as a boxplot. The maximum, third quartile,median, first quartile and minimum of data are showing on the plot. From this graph, there may be a outlier point in the data in 2016. The residual is the point 5333 in Nov 2016.

Service Learning Revenue in 2015

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
479.5	682.0	1178.0	1147.0	1660.0	1953.0

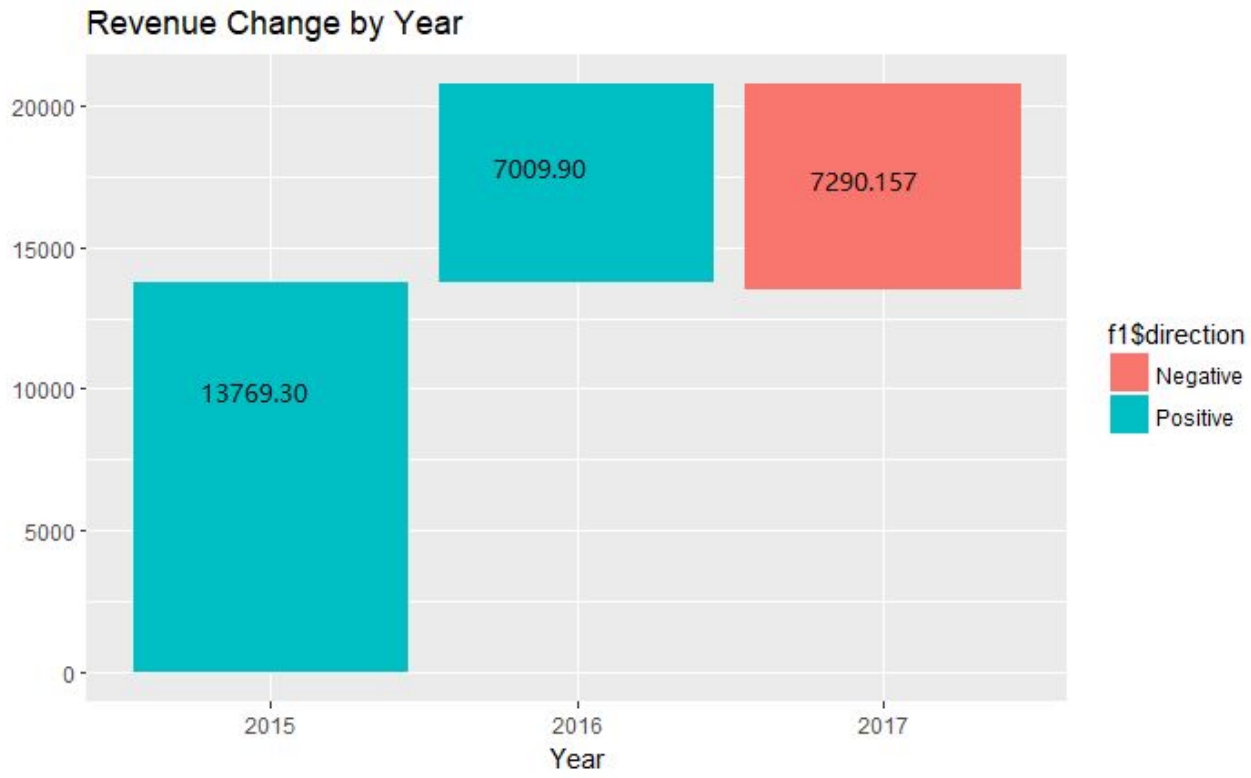
Service Learning Revenue in 2016

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
40.0	885.8	1464.0	1731.0	2110.0	5333.0

Service Learning Revenue in 2017

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
140.0	409.0	991.5	1225.0	1819.0	3346.0

Revenue increase and decrease are compared by years:



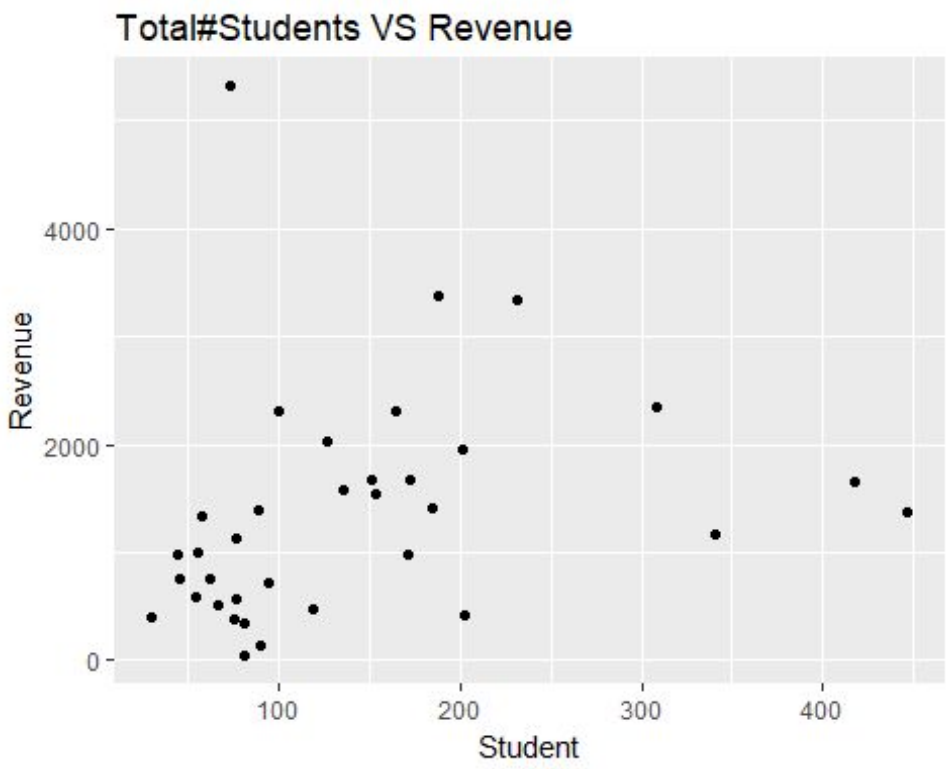
In 2015, the revenue increased $13759.3. From 2016 to 2015, the revenue increased $7009.9. The service learning revenue is $20769.20. However, from 2016 to 2017, the revenue decreased $7290.157. The service learning revenue is $13479.04.

From the data we found that the revenue on 2017 August was 0. Also, the revenues on June and July are much lower than the other months. This phenomenon also appeared in 2015 and 2016. The June, July and August revenues were also lower than the other months in these two years. This might because more of the students are taking summer break.

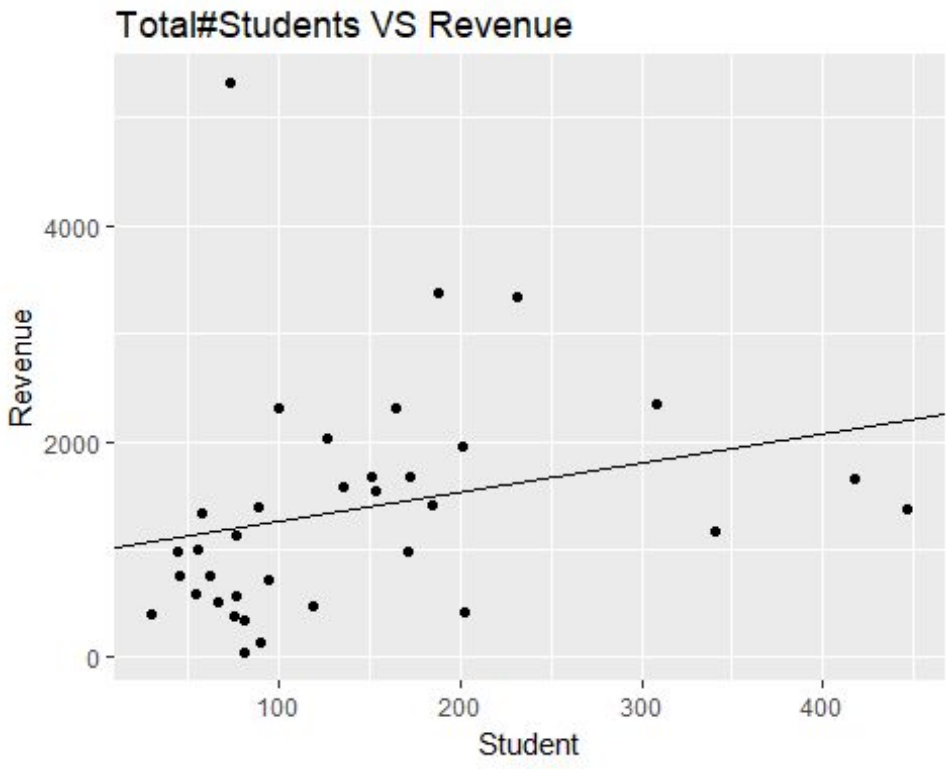
Hypothesis Test:

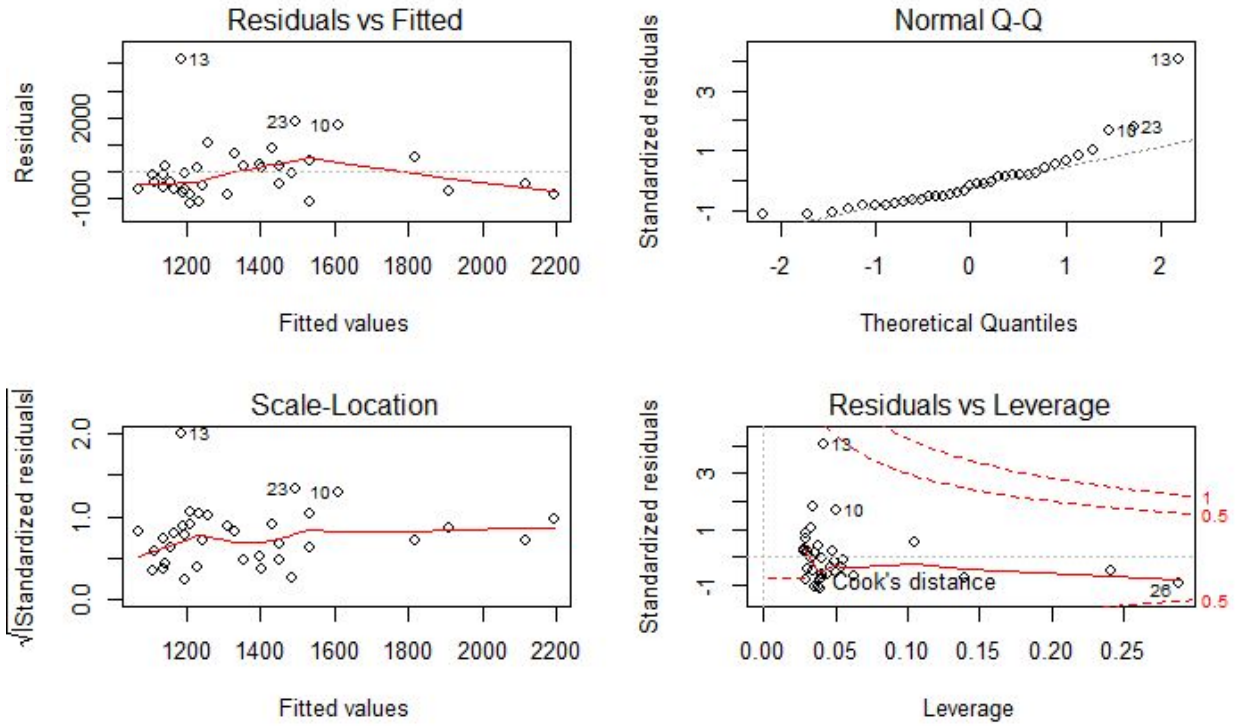
Is the amount of services learning revenue linear related to total number of student?

Can i use total number of student to predict the total amount of services learning?



This graph is showing the relationship of students and revenue. The data of students are independent variables, the data of revenues are dependent variables.





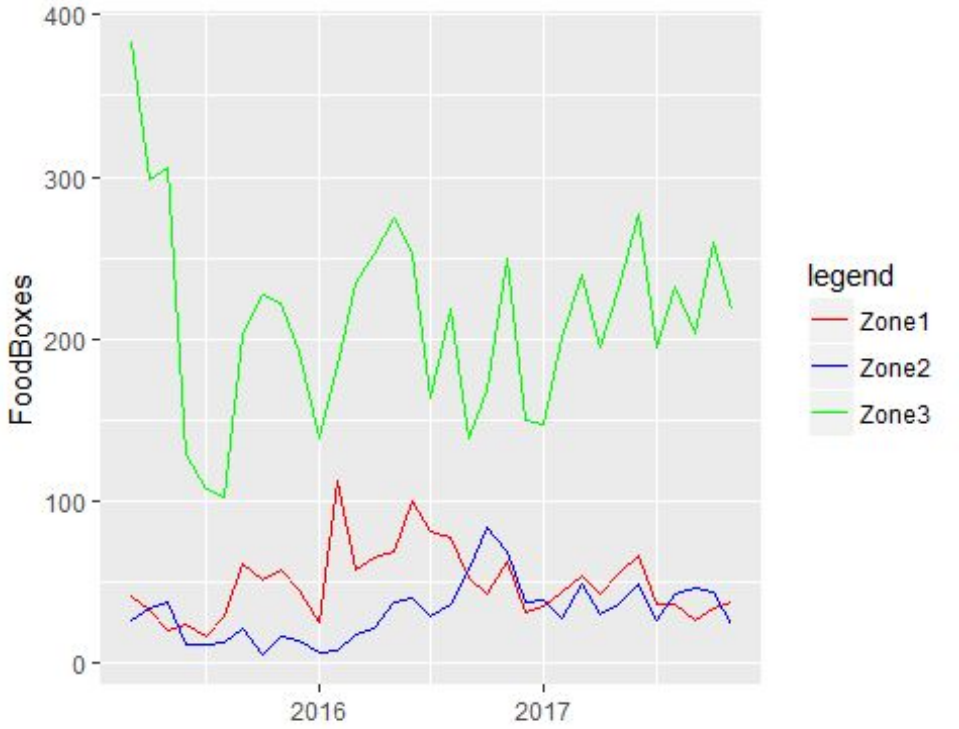
The linear square line is y = 2.686 x + 991.115, the adjusted r square is 0.03. The total number of students and service learning revenue don't have strong linear relationship.

Model assessment:

This qqnorm plot shows the errors are not normal distribution. Based on the graph, the error is not normal distribution. That means the linear regression model is not reliable.

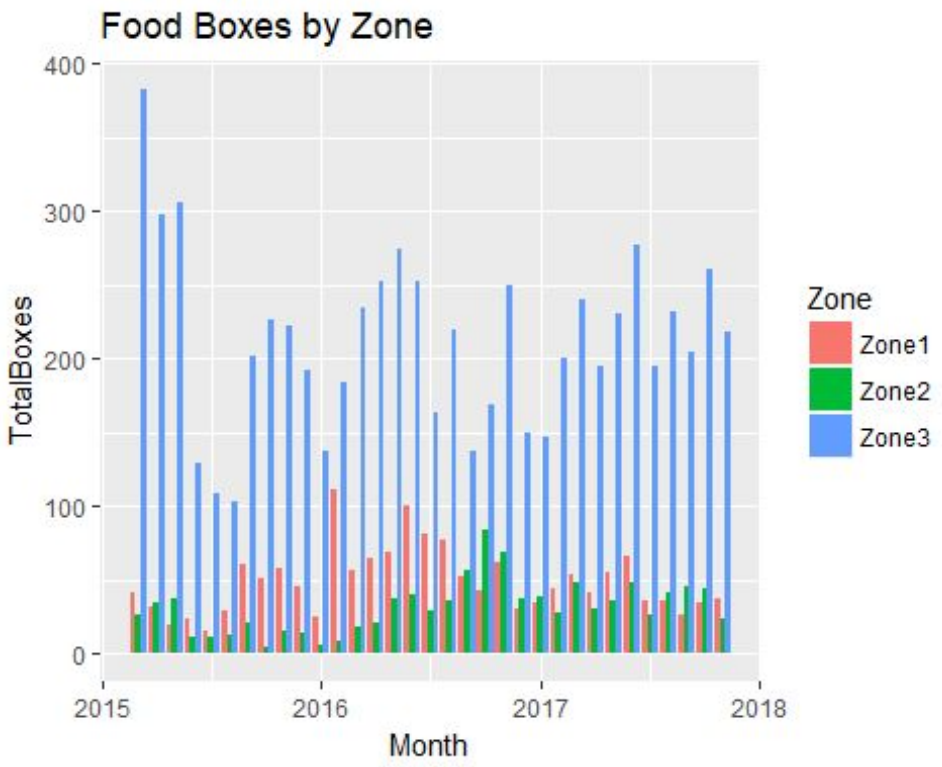
Food Box

Data Visualization/summary:



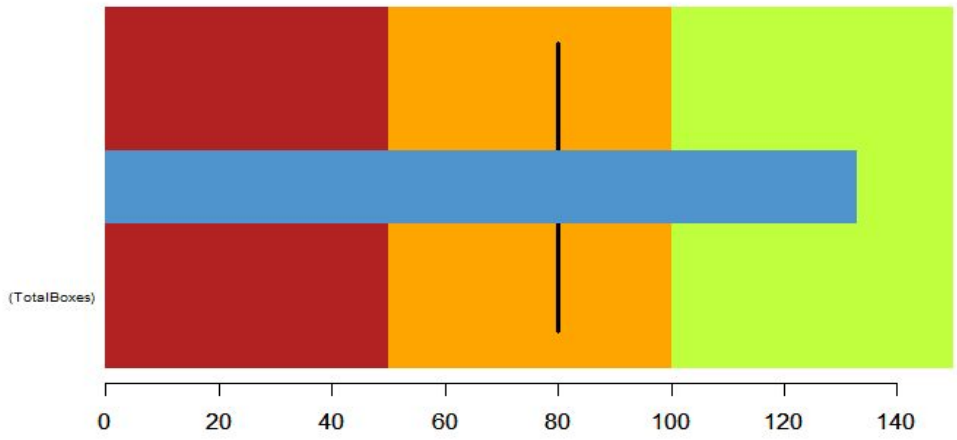
This graph is showing the line graphs of Zone1, Zone2 and Zone3 in the Food Boxes dataset. The data of Zone3 are large than the other two which means there is more demand of food boxes in zone 3. The is might because the GDP of zone 3's residents is higher than the other two.

As the graph shows, the food are more demand one April or May in Zone 3. For Zone 2, the food are more demand on April or September. The food are more demand on February and May in Zone 1. All of these three zones demand less food at the beginning or the end of the year.



This graph describes the food boxes distribution by zone.

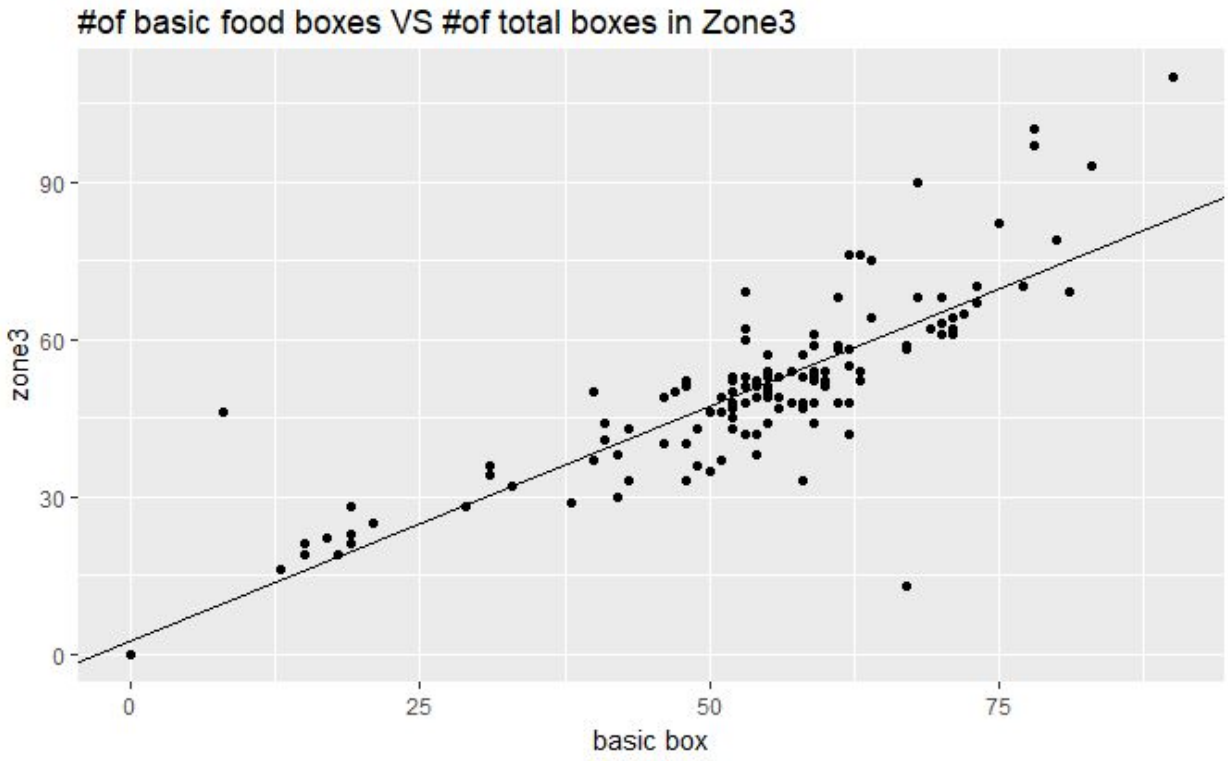
Total Food Boxes



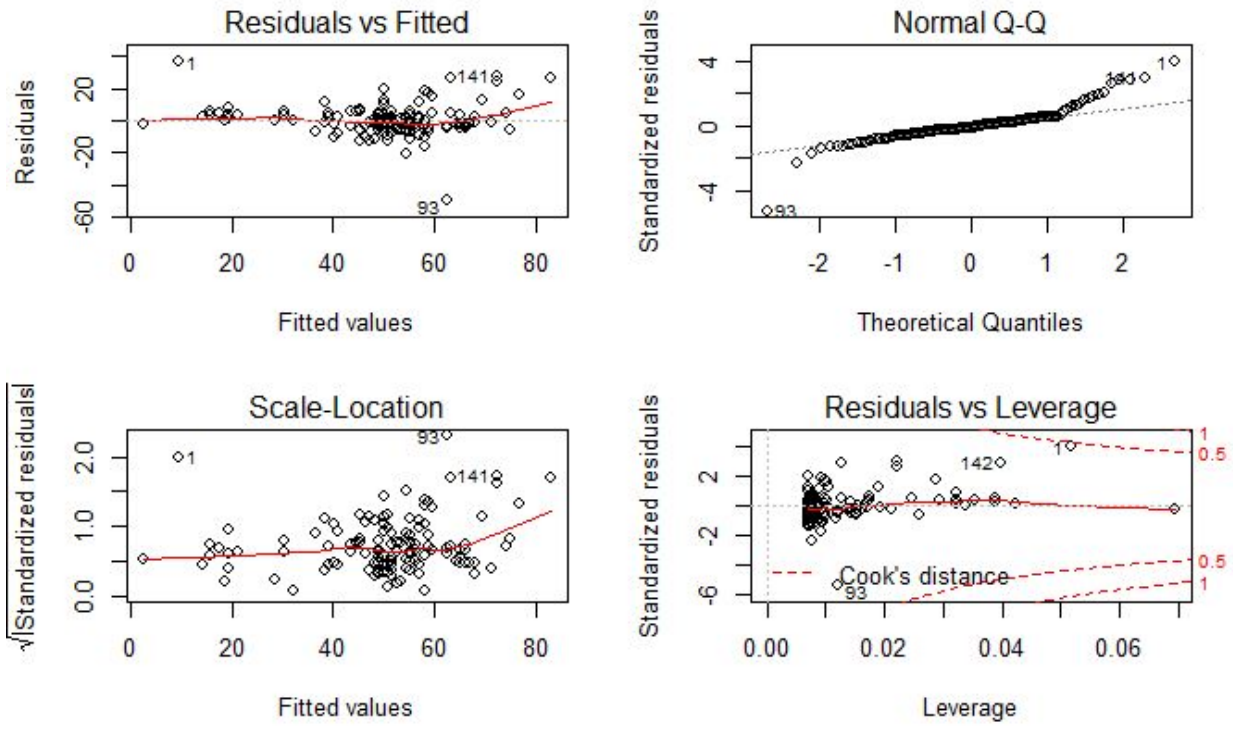
Showing the KPI of total food boxes. The reference point is 80. The three levels of background are 50,100 and 150. The blue bar reach to the green level which is the excellent level. The critical black line is the target value. The target value is 80.

Hypothesis test:

Is zone3 total boxes linearly related to number of basic boxes?



The linear fitted line is y = 0.89559 x + 2.51331. Adjusted R square is 0.7367. The R square is good. the number of family boxes can predict the number of boxes in zone 3 well. All the x variables (the number of basic boxes) can explain 73.67% variation in number of boxes in zone 3.



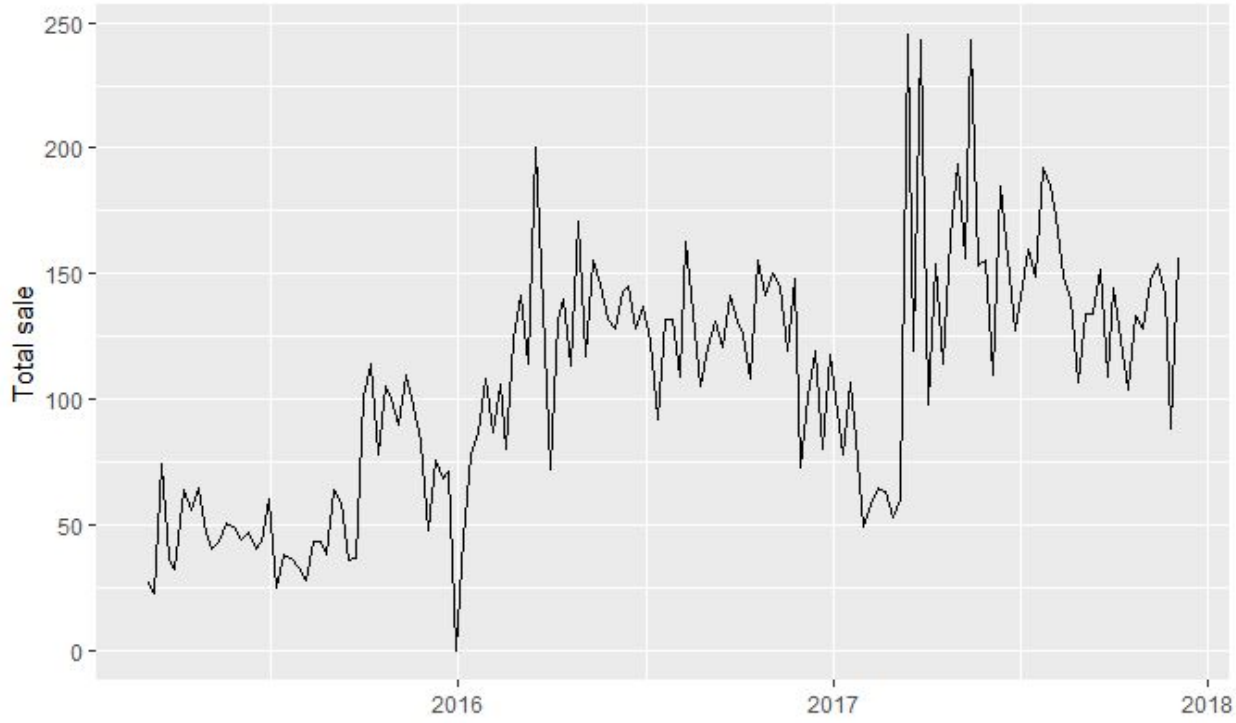
Model assessment:

There is no obvious pattern in the residuals & fitted graph. This means the number of basic boxes has linear relationship with total boxes in zone 3. The Normal Q-Q graph indicates that the residuals are looks normal distribution. Therefore, this is a useful model.

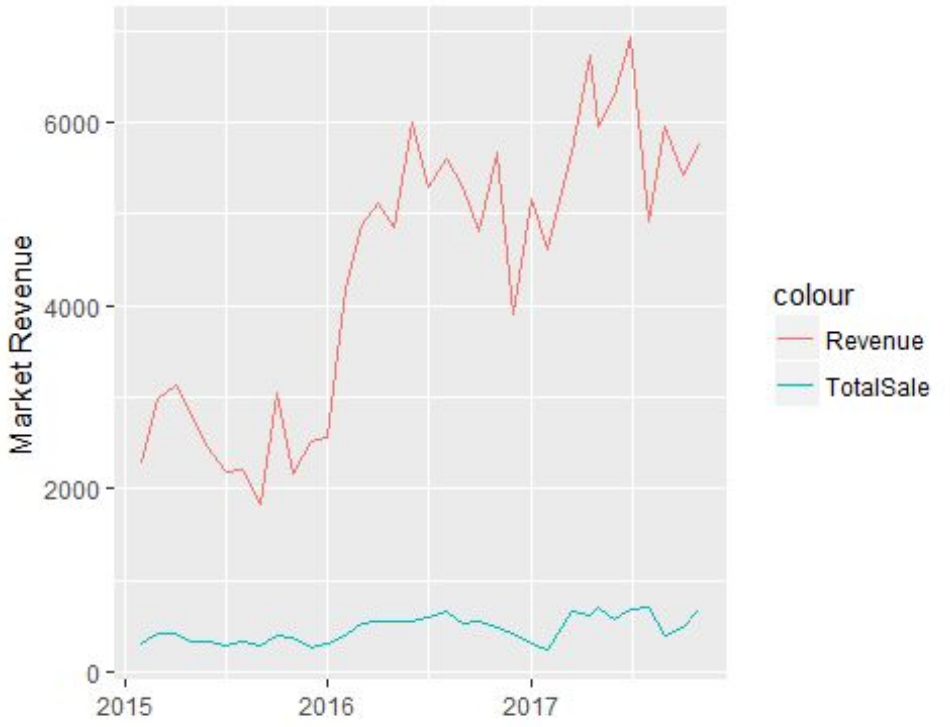
Market

Data Visualization/summary:

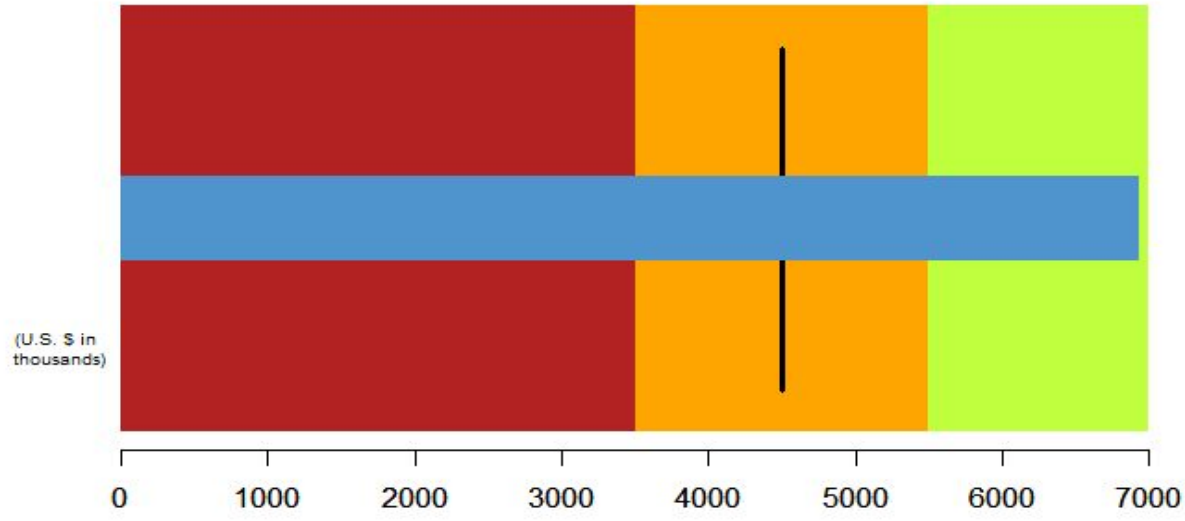
Total Sales:



This graph is showing the total sale from market in different date from data. As the graph shows, the total transactions are growing basically. The overall trend for 3 years is increase.



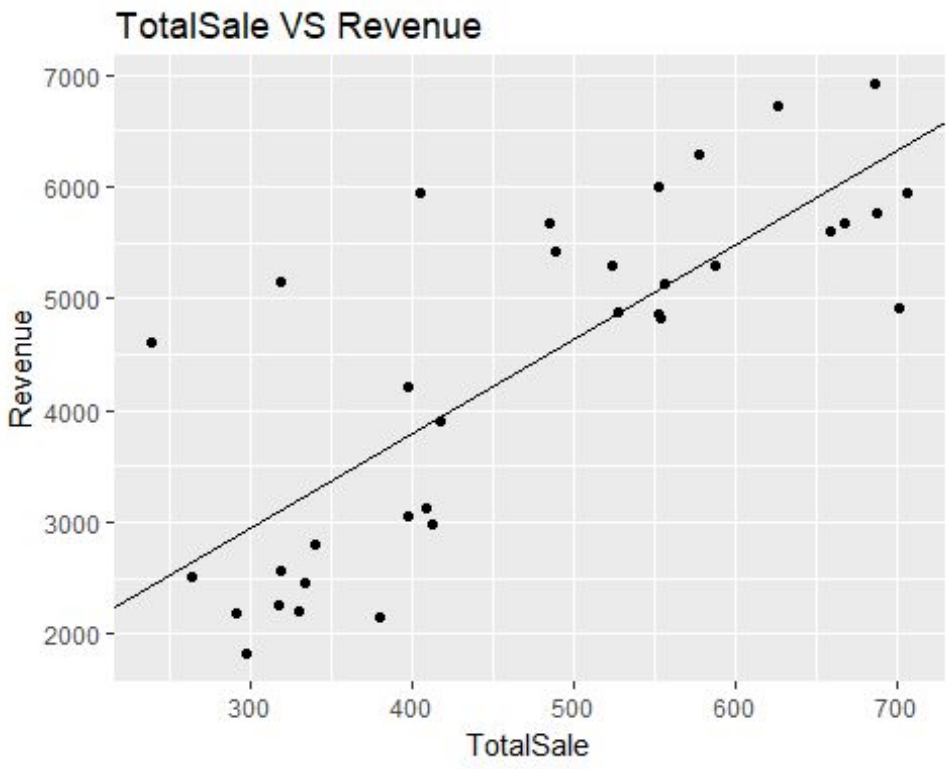
This graph is comparing the Revenue and total sales of market in different data. The total sale is clearly lower than revenue everyday. Since the company is non-profit, so they sale food in a low price. Their revenue might come from people's donation or sponsor's advertisement fee. These cause a high difference between revenues and total sales.



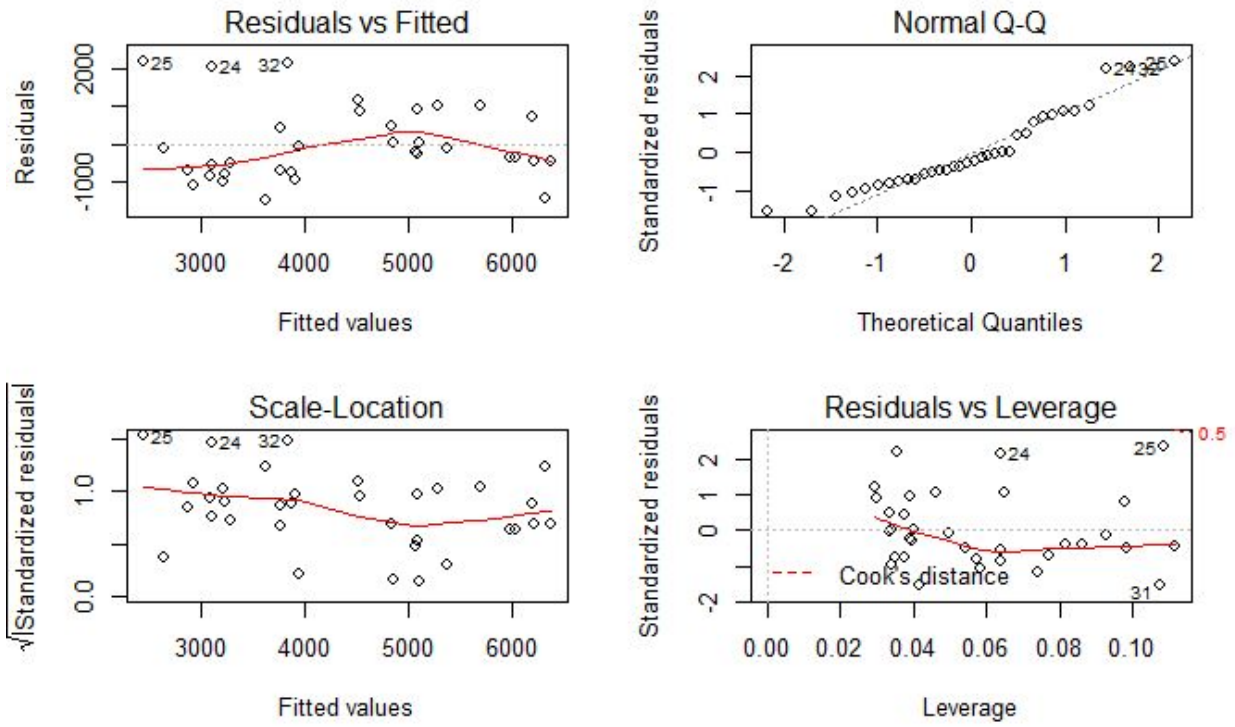
Showing the KPI of market revenue. The reference point of this graph is 4500, the three levels of background are 3500,5500 and 7000. The blue bar reach to the green level which is the excellent level. The critical black line is the target value. The target value is 4500.

Hypothesis test:

is total sale linearly related to revenue?



The linear square fit line : y = 8.431 x + 419.221. Adjusted R square is 0.5983.



Model assessment:

There is no obvious pattern in residuals vs fitted graph. The relation between total sales and revenue are linear. From the Normal Q-Q graph, the residuals are not normal distribution. Therefore, the model is reliable.

Results and Conclusion

We primarily explore 3 parts of the entire data: Service Learning Revenue, Food Box and Market

For the service learning revenue part, we found that there is an outlier in the service learning part. In Nov 2016, there is a abnormal service learning revenue point. In that point, the revenue reach to 5333 which is especially high. The other interesting results we found from this part is the service learning revenue increases about $7000 from 2016 to 2015 and falls down $7200 from 2017 to 2016. Also, the company could focus on how to increase education revenue during summer.

For the food box part, the company delivers two kind of boxes (basic boxes and family boxes) to 3 zones(zone1, zone2 and zone3). We found that the basic box and the number of boxes delivered to zone 3 have linear relationship. The most food boxes delivered to zone 3 is basic box. Company can use the number of basic box to predict how many food boxes should be delivered to zone 3.

The last part is market. We found that revenue is much higher than company's market sales. From their website, we found that in 2017, almost 75% of their companies' income was from contribution. Therefore, we think the reason why revenue is higher than sales is donation. Also, the trend of market sales is increasing.

Feature Engineering:

- KPI analysis
- Waterfall graph to show revenue change by year
- Linear regression analysis