GOODBELLY:

Using Statistics to Justify the Marketing Expense

by Shuangjiao Hu, Shenyi Lu
03/22/2018

CS 686: Machine Learning

UNIVERSITY OF
SAN FRANCISCO
CHANGE THE WORLD FROM HERE

1 Executive Summary

Marty Wellbing, GoodBelly's marketing manager, want to analysis the selling data to justify the demo and endcap activities in GoodBelly's promotional program.

We ran backward elimination algorithm on Interaction Regression Models on revenue and sale units with each promotional programs to find the program that is most closely related to Weekly Sales or Revenue.

After executing backward elimination on several interaction regression models and multiple linear regression models, we advised Marty to use regional sales representative and endcap promotion together, and run in-store demonstrations program once per store. According to our final model, the demos do boost sales, and sales would not revert to normal levels shortly afterwards. Regional sales representative and endcap promotion are not effective when used individually, but would boost sale significantly when used together in one store.

2 Problem Statement

NextFoods Inc. produced a new line of probiotic juice products, GoodBelly, and had the challenge of raising product awareness. Due to the limited budget, GoodBelly allocated much of its small marketing budget to three promotional programs, including in-store demonstrations program, regional sales representative and endcap promotion. Then management was pressured to cut any marketing expense that did not directly contribute to GoodBelly's results, and use statistical analysis to justify their decision.

3 Data Analysis

3.1 Data Description

The data provided by the representatives includes 12 columns.

0. Date: The date the representatives recorded the data
1. Region: The state that the store located
2. Store: The store that sells GoodBelly products.
3. Weekly Sales (Volume): The number of units sold per store per week.
4. Average Retail Price: The average retail price for GoodBelly products per store per week.
5. Sales Rep: Defined as 1 if the store had a regional sales rep (face-to-face contact) and 0 if the store had only the national sales rep (no face-to-face contact).
6. Endcap: Defined as 1 if a store participated in an endcap promotion.
7. Demo: Defined as 1 if the store had a demo on the corresponding week.
8. Demo1-3: Defined as 1 if the store had a demo 1-3 weeks ago.
9. Demo4-5: Defined as 1 if the store had a demo at least 4-5 weeks ago.
10. Natural Retailers: The number of other natural retailers within 5 miles of each store.
11. Fitness Centers: The number of fitness centers within 5 miles of each store.

3.2 Dependent Variable

Since our study is focusing on marketing strategies, we are going to choose sales-related variables as the dependent variable. Since raw data provides Weekly Sales and Average Retail Price, we calculated average revenue of sales per week, Revenue, as a possible dependent variable.

Thus, we can choose Weekly Sales or Revenue as our dependent variable. We built multiple models based on one of the dependent variables, and tried to find a model with highest quality of prediction.

3.3 Independent Variables

Project requirement shows that some predictor variables will not be considered as related to dependent variable. Thus, we dropped Date, Region and Store from our model. Result of backward elimination shows that Natural and Fitness's p-value do not meet alpha = 0.05 threshold, and should be dropped.

4 Methods and Results Analysis

4.1 Model building strategy

Since all independent variables were binary dummy variable, we have decided to use multiple linear regression with interaction regression model to establish a clear statistical relationship between Weekly Sales/Revenue and possible independent variables in data that Wellbeing received for the given 10 weeks.

Our major challenge in model building process is that, as all independent variables are binary dummy variable, Simple Linear Regression model and several other common data analysis methods can not describe the statistical relationship between dependent and independent variables correctly. Polynomial regression models will not be effective in this case, as binary dummy variables always equal 0 or 1 in different orders.

We have ran a backward elimination algorithm that established a regression model that contains all available independent variables in the model, and eliminate them sequentially, until the p-values of remaining variables are all below alpha = 0.05. The regression model will use either Weekly sales or Revenue as dependent variable against different combinations of interaction terms.

By running backward elimination algorithm on general additive model, Natural Retailers and Fitness Centers are dropped from possible independent variables as they do not meet threshold in the process.

4.2 MLR model selection

4.2.1 Possible models

By running backward elimination algorithm on a regression model that contains all available independent variables in the model, and choose Weekly Sales or Revenue as dependent variable, we can get two similar simple additive model.

By running backward elimination algorithm on several interaction regression model based on simple additive model, we selected two similar interaction regression model among results that would likely describe the statistical relationship between independent and dependent variables. Since Price is highly correlated with Revenue, we will drop Price in models that use Revenue as dependent variable.

To simplify the equation, we will use certain abbreviation.
Weekly Sales: Unit; Sales Rep: Rep

Model 1: Simple additive model

$$Unit \sim Rep + Endcap + Price + Demo + Demo_{1\_3} + Demo_{4\_5}$$
$$Revenue \sim Rep + Endcap + Demo + Demo_{1\_3} + Demo_{4\_5}$$

Model 2: Interaction regression model

$$Unit \sim Rep * Endcap + Price + Demo * Demo_{1\_3} * Demo_{4\_5}$$
$$Revenue \sim Rep * Endcap + Demo * Demo_{1\_3} * Demo_{4\_5}$$

Backward elimination result:

$$Unit \sim Rep * Endcap + Price + Demo + Demo_{1\_3} + Demo_{4\_5}$$
$$Revenue \sim Rep * Endcap + Demo + Demo_{1\_3} + Demo_{4\_5}$$

4.2.2 Analysis of possible model

OLS Regression result:

Simple additive model 1:

$$Unit = 77.0Rep + 305.0Endcap - 28.6Price + 111.3Demo + 73.7Demo_{1\_3} + 67.7Demo_{4\_5} + 294.2$$
$$R^2 = 0.67, \quad R_a^2 = 0.67$$

Simple additive model 2:

$$Revenue = 353.4Rep + 1159.6Endcap + 470.5Demo + 279.5Demo_{1\_3} + 303.7Demo_{4\_5} + 708.6$$
$$R^2 = 0.65, \quad R_a^2 = 0.65$$

Interaction regression model 1:

$$Unit = 59.5Rep + 0.6Endcap + 453.8Rep * Endcap - 22.1Price + 106.8Demo + 73.4Demo_{1\_3} + 74.6Demo_{4\_5} + 276.6$$
$$R^2 = 0.81, \quad R_a^2 = 0.81$$

Interaction regression model 2:

$$Revenue = 294.6Rep - 9.9Endcap + 1736.9Rep * Endcap + 453.7Demo + 275.4Demo_{1\_3} + 331.8Demo_{4\_5} + 740.0$$
$$R^2 = 0.77, \quad R_a^2 = 0.77$$

4.2.3 Final model

Based on OLS regression result, we choose Interaction regression model 1 as our final model, as this model has highest coefficient strength. As independent variables are not highly interactive, we choose MLR Type II ANOVA Table for ANOVA analysis. (Please refer to appendix for ANOVA Table and model performance.)

4.3 Final model interpretation

Our final regression model for Weekly Sales is

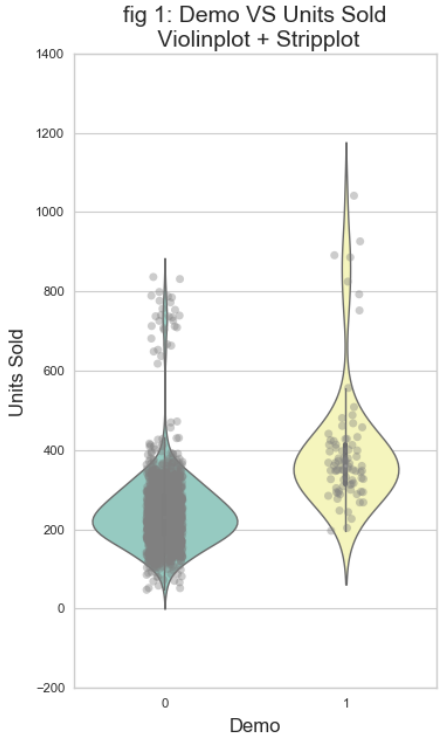
$$Unit = 59.5Rep + 0.6Endcap + 453.8Rep * Endcap - 22.1Price + 106.8Demo + 73.4Demo_{1\_3} + 74.6Demo_{4\_5} + 276.6$$

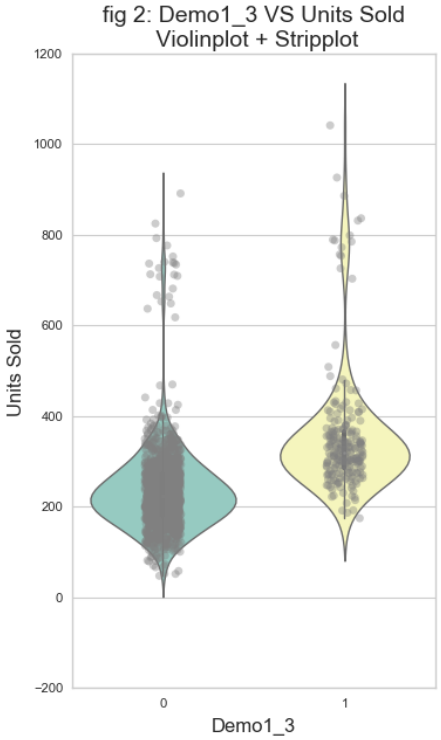
$$R^2 = 0.81, \quad R_a^2 = 0.81$$

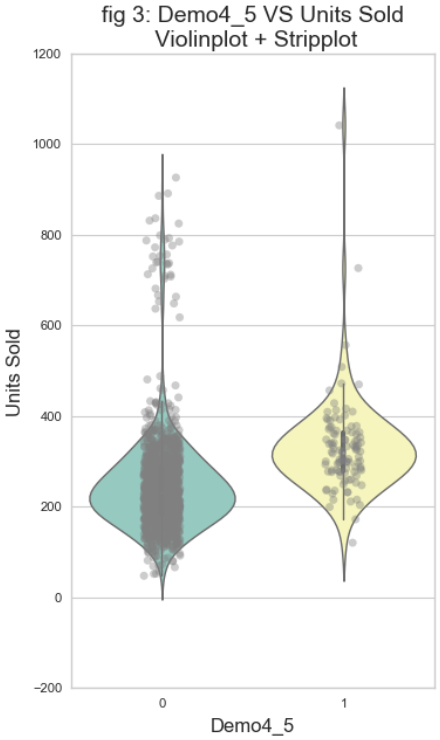
4.3.1 Demo

Demo, Demo1_3 and Demo4_5 have significant positive coefficients, and the coefficients of Demo1_3 and Demo4_5 is about 70% of coefficient of Demo. Which means the effect of in-store demonstrations would not revert in short time. Fig 1 - fig 3 also shows significant increase in sales, which can also prove that the result of the effect of in-store demonstrations can last for weeks.

But the interaction terms of Demos do not meet threshold in backward elimination. The lack of interaction terms could mean that holding in-store demonstrations repeatedly in short terms do not further boost sale, and we do not have sufficient data to show that repeatedly holding demonstrations in mid to long term could maintain the boost.





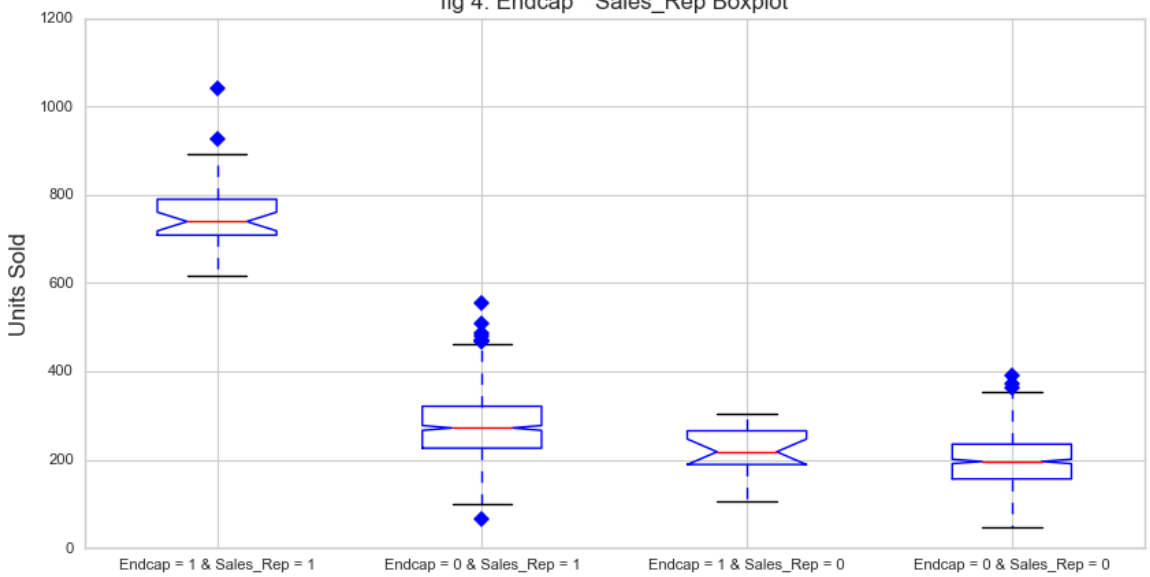


4.3.2 Sales Rep and Endcap

Rep's coefficient is positive but smaller than all three Demos, Endcap's coefficient is positive but close to zero. The interaction term of Rep and Endcap have a large positive coefficient. It is possible that holding regional Sales Rep and Endcap in the same store would significantly increase sale, but holding them individually could not be very effective.

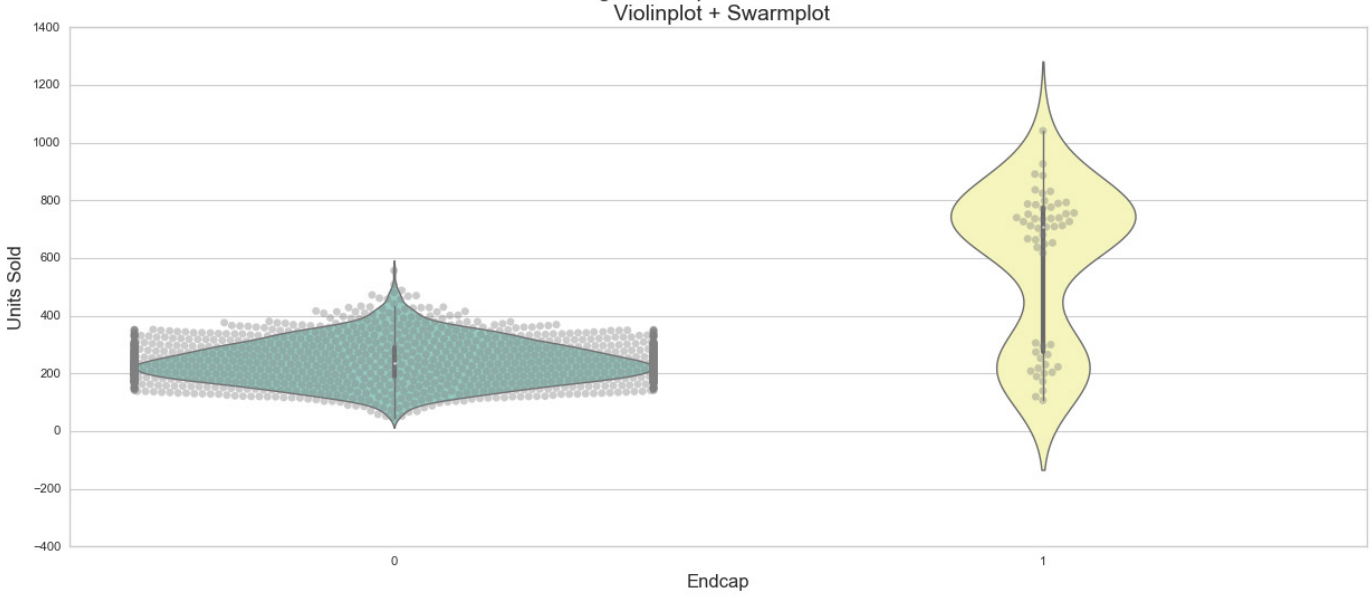
We drew a boxplot that have Endcap and Sales Rep combined as our X axis, and Units Sold as Y axis. We can see a significant increase when both of them equal to 1 while only one of them equals to 1, the increase is relatively slight.

fig 4: Endcap * Sales_Rep Boxplot



It is worth noticing that out of 1386 stores, 762 stores hold Sales Rep, only 53 stores hold Endcap, out of which only 36 stores hold both programs. In this case, only holding Sales Rep is less likely to increase sales and we may need further research. Also, from Fig 5 we can clearly find the unusual distribution of data when Endcap equals to 1. Although in the following passage, we have given an explaination for this situation, this might simply because of lack of data.

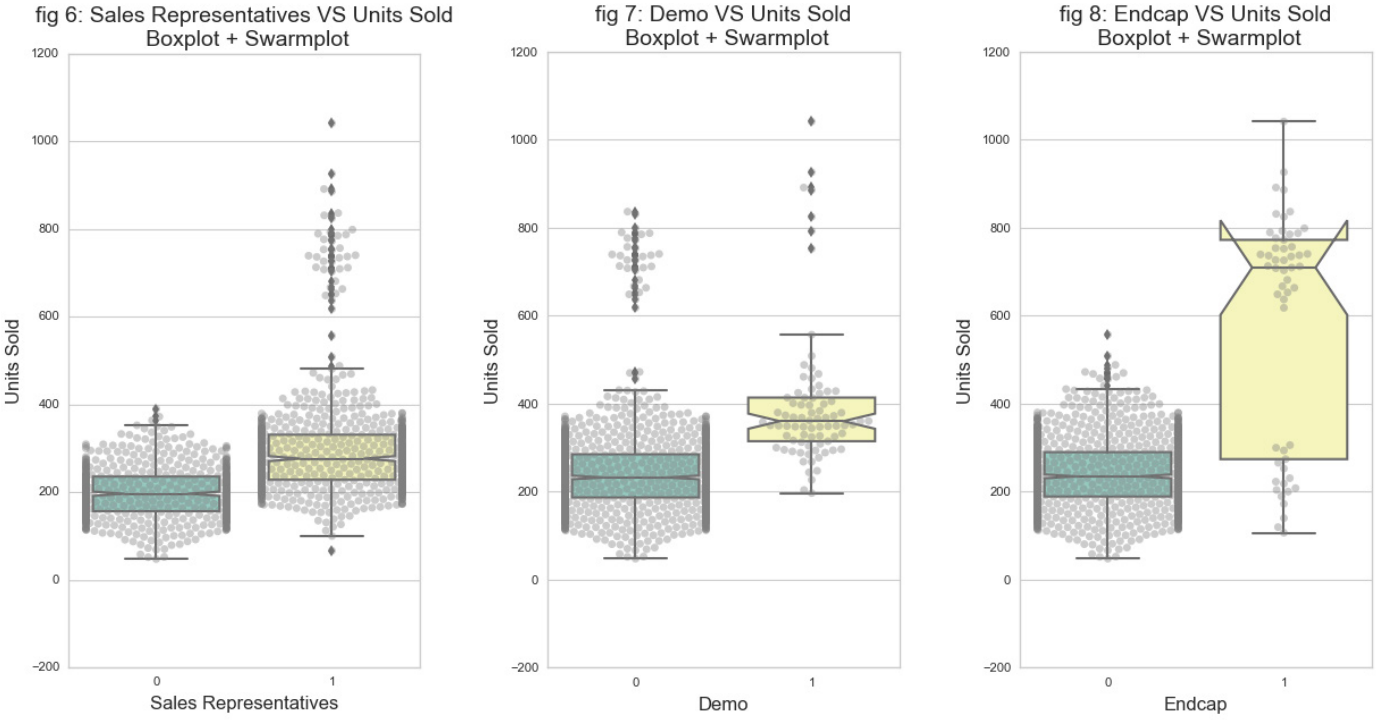
fig 5 Endcap VS Units Sold
Violinplot + Swarmplot

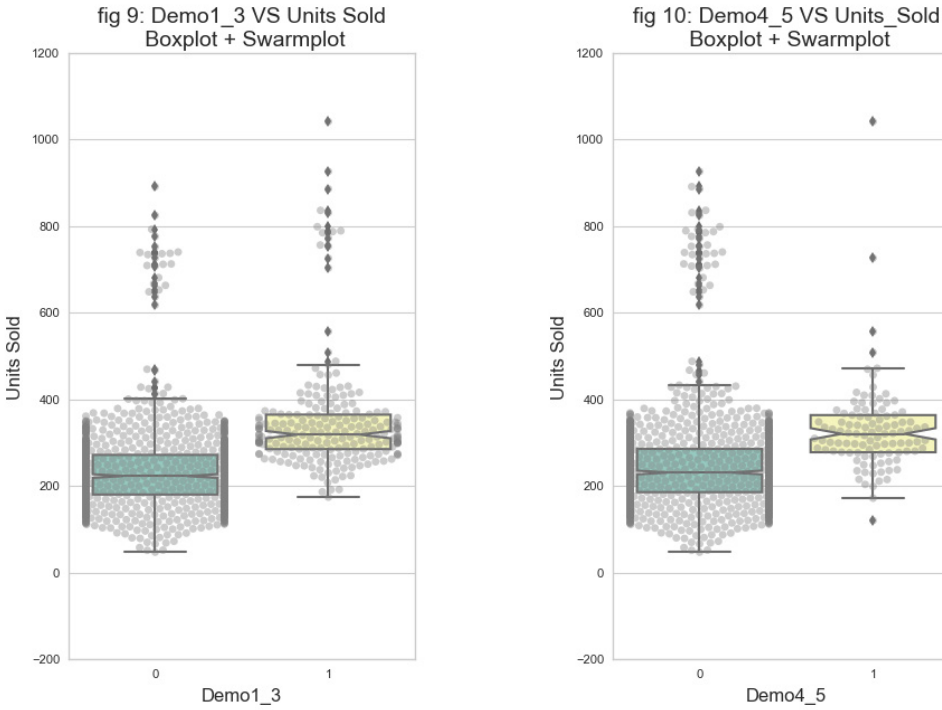


4.4 Graph Analysis & Model Test

Hypothesis test and many other data analysis tools are not effective, because most independent variables in this research are binary. Therefore, we choose box plots, violin plots and swarm plots to make a visualization of data, for us to have a intuitive thought.

Fig 6 - 10 shows that each promotional program is related to significant increase of the mean of dependent variables.





However, on the boxplot and swarmplot of Endcap, when a store participated in an Endcap promotion. (Endcap = 1), the data points were unusually separated into two groups. The plots also show a large amount of outliers. In order to analyze it, in fig 11 - 16, swarmplots divided data into two groups by colors.

In fig 11, when Sales Rep equals to 1, it is clear that the outliers of the data is caused by reinforcement effect. In fig 12, when Endcap equals 1, the two separeted data group also came from reinforcement effect caused by Sales Rep. From fig 12, we see that even if there was no Endcap, the Sales Rep was also strongly related to the increase of Units Sold but Endcap without Sales Rep may contribute very little to Units Sold.

All these situation we observed from plots matches our model that the coefficient of interaction term is very high and coefficient of Sales Rep is much higher than Endcap.

fig 11: Sale Representatives VS Units Sold
Scatterplot(color divided by Endcap)

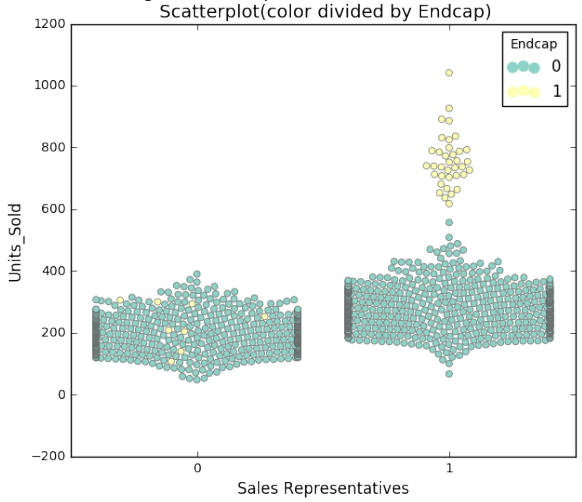
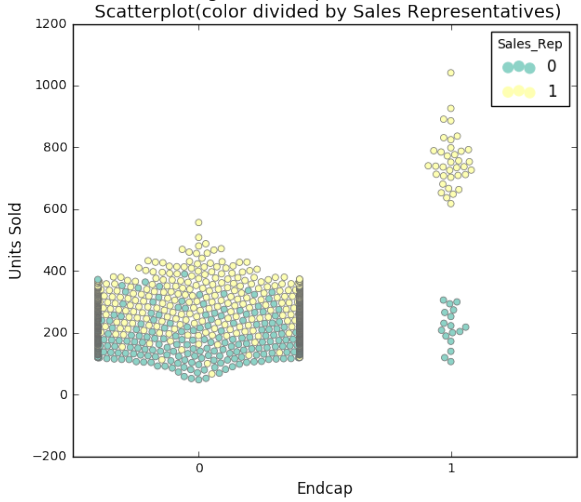


fig 12: Endcap VS Units Sold
Scatterplot(color divided by Sales Representatives)



In fig 13 and 14, for Demo, the outliers are also caused by reinforcement effect. However, no matter Demo equals to 1 or 0, the reinforcement effect keeps the same, which proves our model that the coefficient of interaction term between Demo and Endcap or Demo and Sales Rep is 0.

fig 13: Demo VS Units_Sold
Scatterplot(color divided by Endcap)

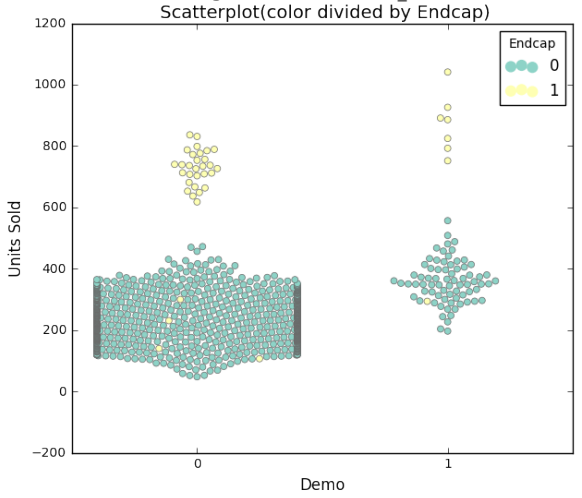
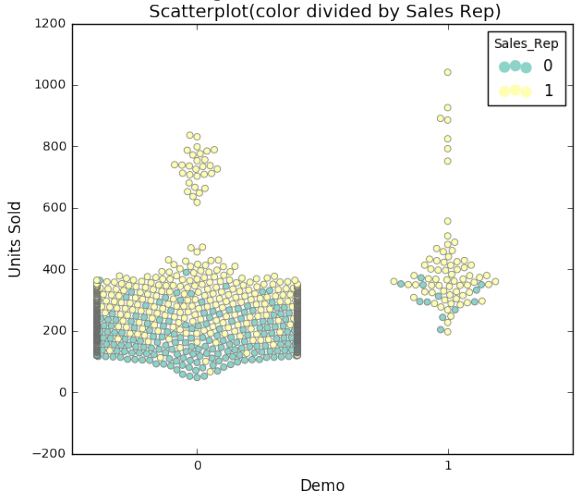


fig 14: Demo VS Units Sold
Scatterplot(color divided by Sales Rep)



From the plots we made, we can basically conclude that our final model is highly related to the data provided by sales representatives. Most of the information the model convey to us can be found in the plots of data, so the model should be reliable.

To see more plots related to the outliers, reinforcement effect and increase of revenue or units sold, please also refer to the appendix.

5 Conclusion

Based on the models we made and other analysis, we have made the conclusion that all three marketing strategy are effective, so we should not cancel them. There's a relatively simple statistical relationship between Weekly Sales against Sales Rep, Endcap and Demos.

Final model shows that all three marketing strategies are effective in boosting sales, and there are interaction between different marketing strategies that can be exploited to boost sales further.

It seems quite likely that in-store demonstrations program boost sales significantly, and this effect can last for weeks. Therefore, it might be pointless to do demostration too frequently.

While holding in the same store, regional Sales Rep and Endcap could boost sales significantly, but much less effective otherwise. Based on our model and plots, holding Endcap without Sales Rep almost have no contribution to sales, while having Sales Rep can still slightly increase sales.

Due to lack of cost-related data, we cannot determine whether the increase in sales volume could justify the associated costs. The small number of stores that holding Endcap also worth further notice, to determine whether there are other factors contribute to the strong reinforcement effect when regional Sales Rep and Endcap are held together.

6 Appendix

The appendix includes some other works we did that highly related to our case study. Including a complete plots for revenue analysis, some supplementary plots for analysis of Sales Units, ANOVA table and model performance summary of our final model.

6.1 Data exploration

The data exploration part is divided into two parts, one is about revenue and the other one is about Units Sold. The division is based on the dependent variables we selected to build the model. The revenue is more related to actual sales, although it is not the dependent variable for our final model. Therefore, if the manager is also interested in the weekly revenue, he can also find a complete analysis plots that associated with the revenue in the following content.

6.1.1 Revenue as dependent variable:

6.1.1.1 Box plots and swarm plots based on revenue

fig 15: Sale Representatives VS Revenue
Boxplot + Swarmplot

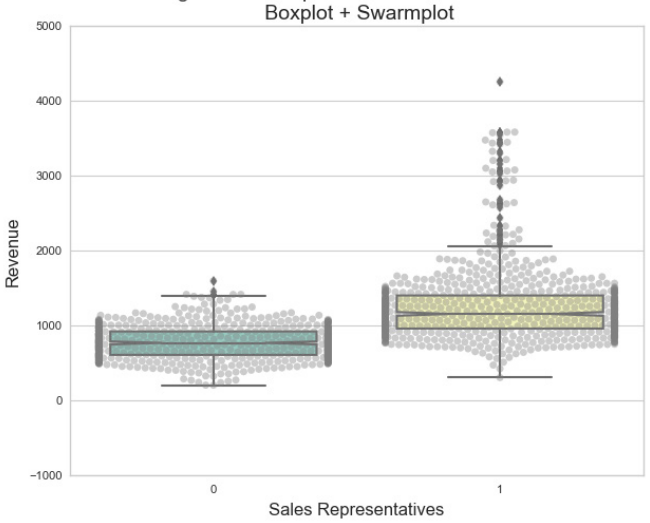


fig 16: Endcap VS Revenue
Boxplot + Swarmplot

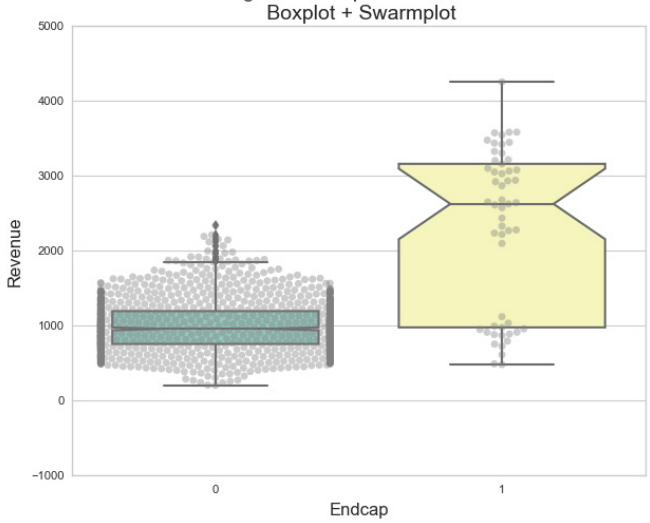


fig 17: Demo VS Revenue
Boxplot + Swarmplot

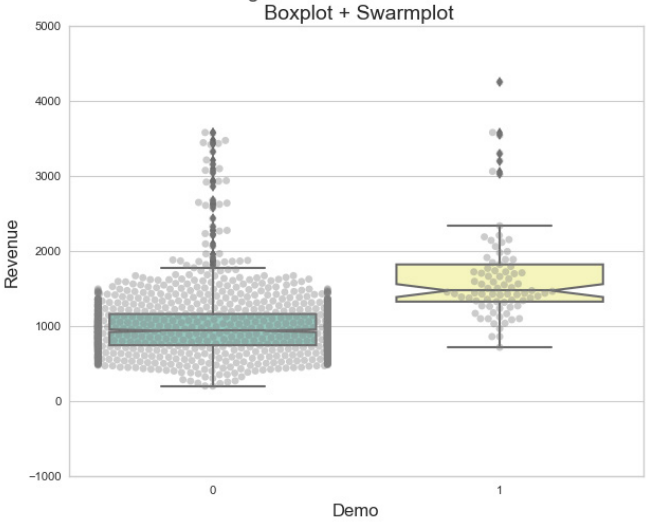


fig 18: Demo1_3 VS Revenue
Boxplot + Swarmplot

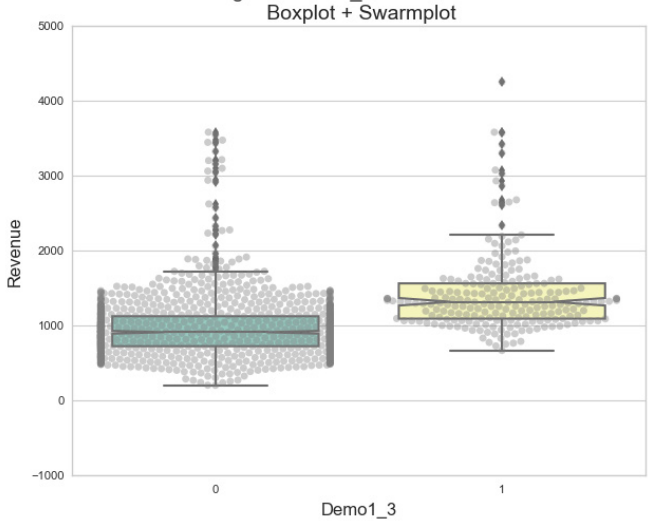
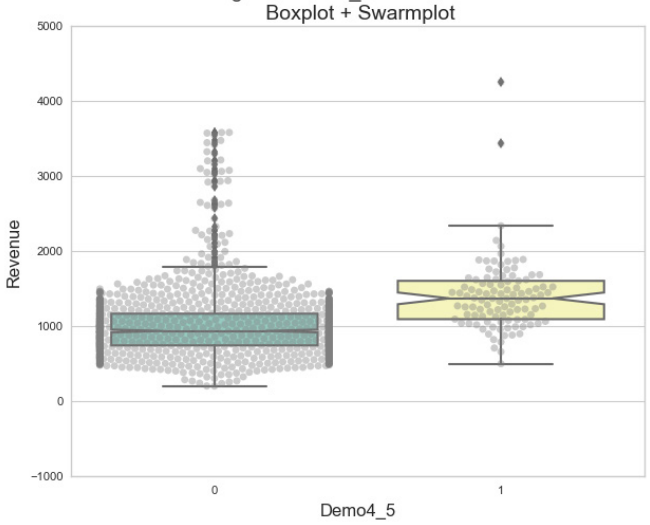


fig 19: Demo4_5 VS Revenue
Boxplot + Swarmplot



6.1.1.2 Violin plots and strip plots based on revenue

fig 20: Sale Representatives VS Revenue
Violinplot + Stripplot

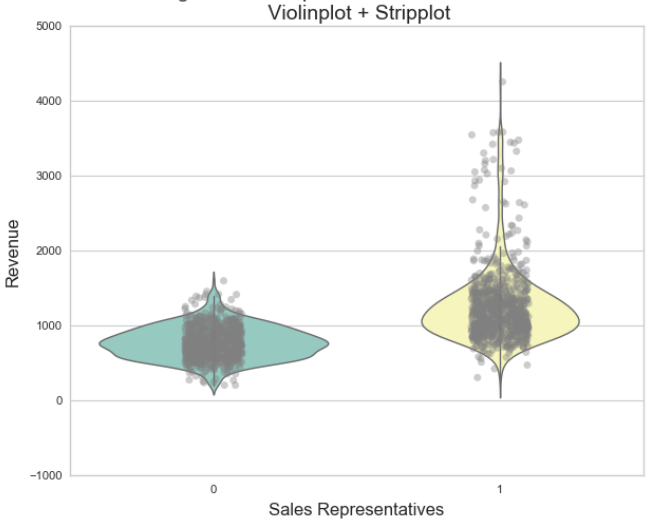


fig 21: Endcap VS Revenue
Violinplot + Stripplot

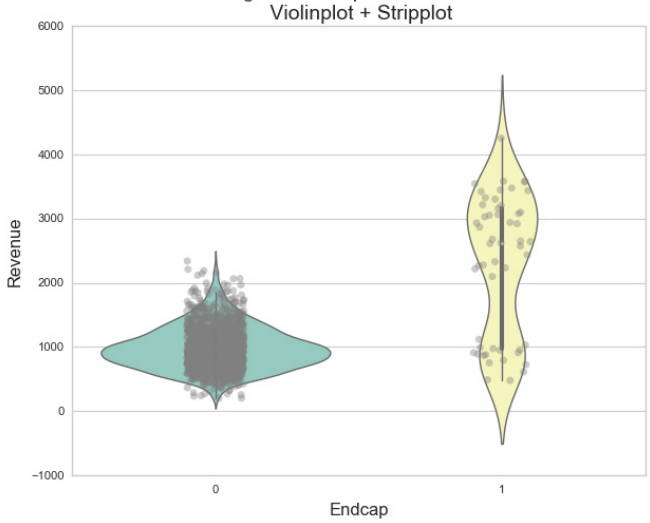


fig 22: Demo VS Revenue
Violinplot + Stripplot

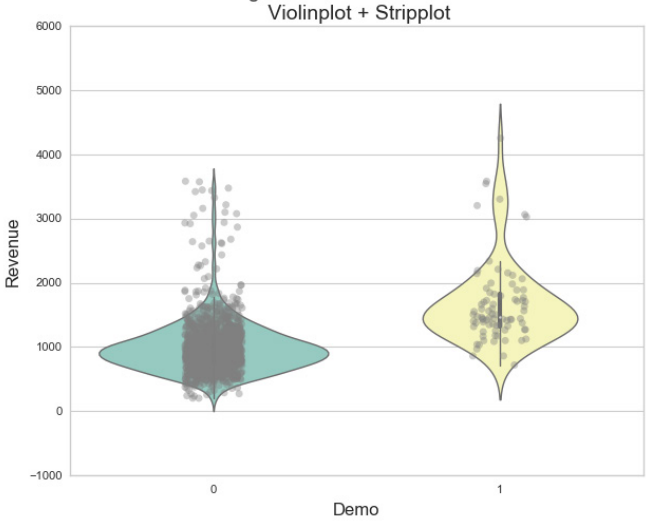


fig 23: Demo1_3 VS Revenue
Violinplot + Stripplot

fig 24: Demo4_5 VS Revenue
Violinplot + Stripplot

6.1.1.3 Swarm plots of sale representatives divided into groups based on revenue

fig 25: Sale Representatives VS Revenue
Scatterplot(color divided by Endcap)

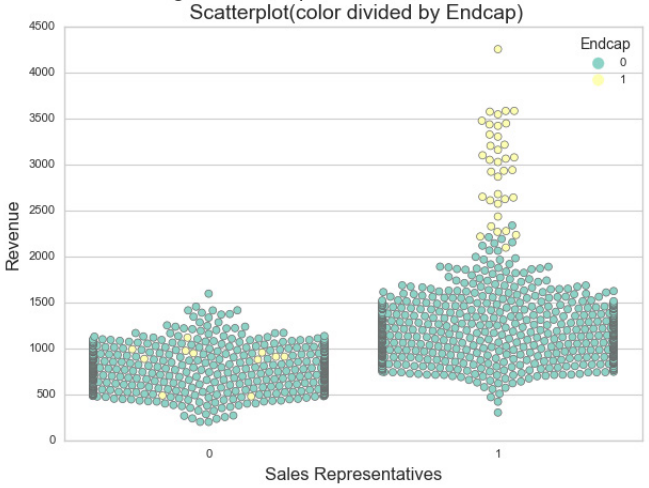


fig 26: Sale Representatives VS Revenue
Scatterplot(color divided by Demo)

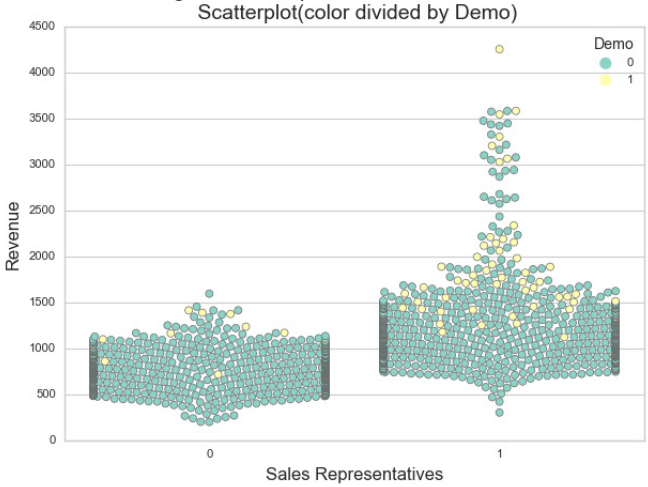


fig 27: Sale Representatives VS Revenue
Scatterplot(color divided by Demo1_3)

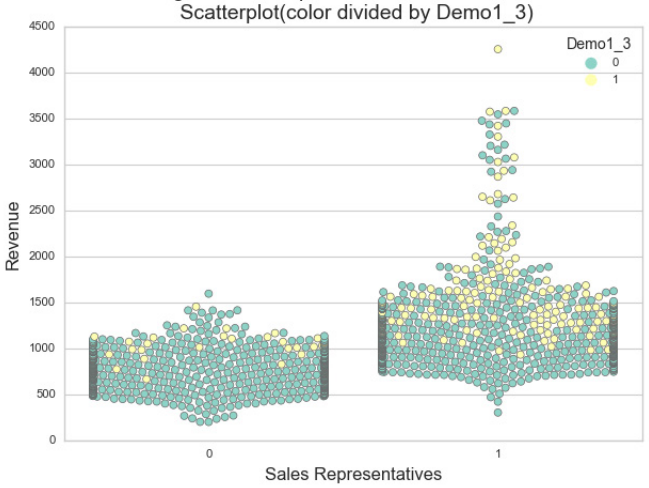
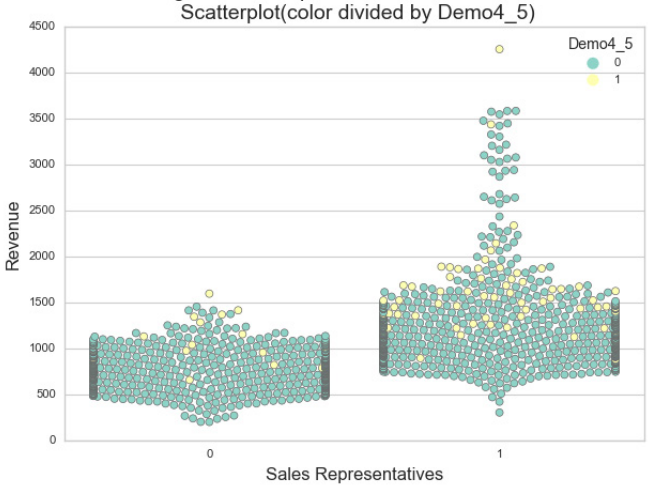


fig 28: Sale Representatives VS Revenue
Scatterplot(color divided by Demo4_5)



6.1.1.4 Swarm plots of endcap divided into groups based on revenue

fig 29: Endcap VS Revenue
Scatterplot(color divided by Sales Representatives)

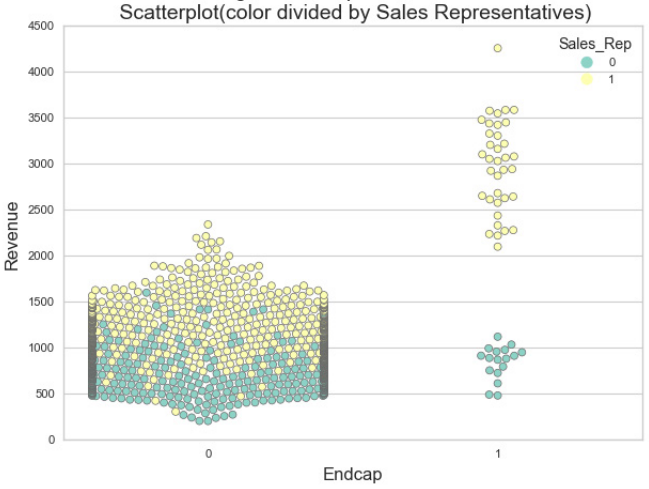
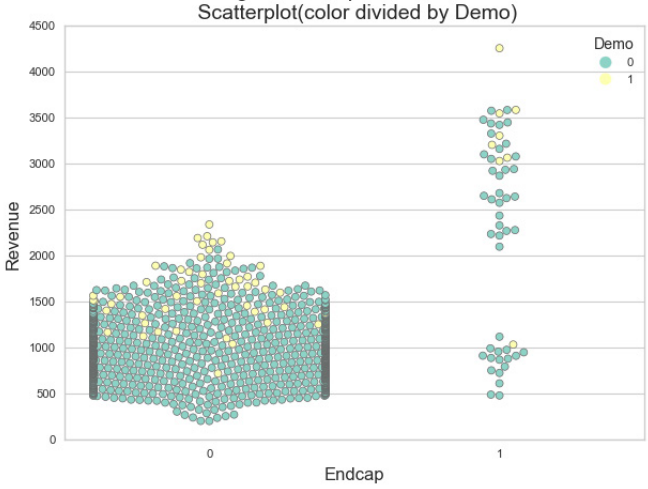
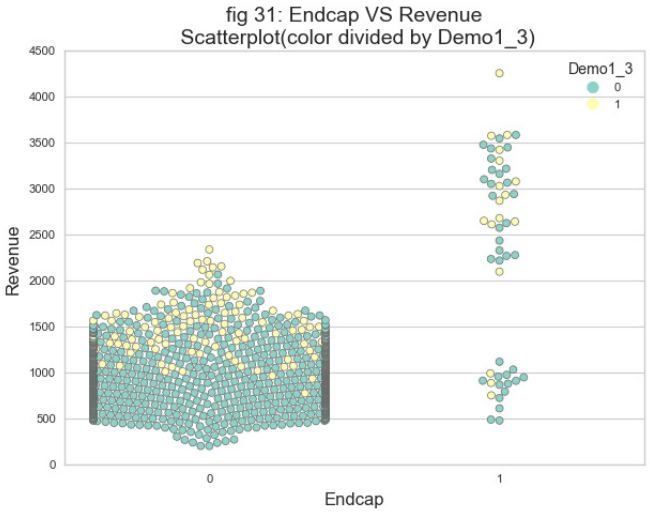
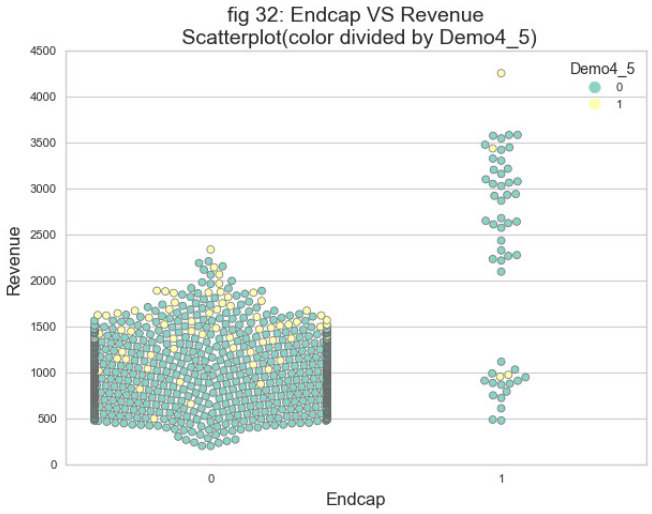


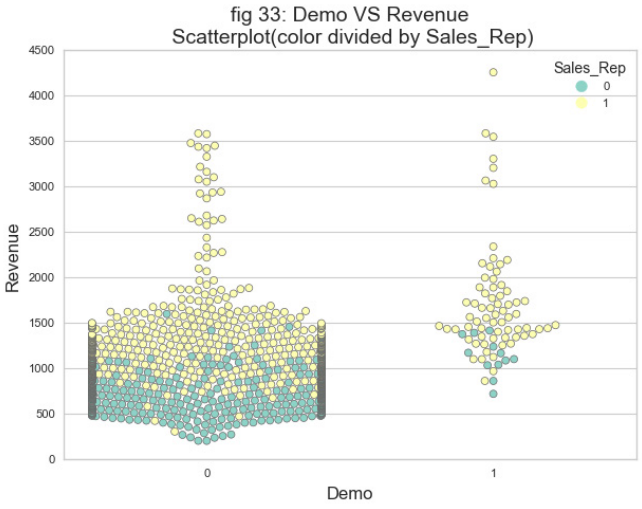
fig 30: Endcap VS Revenue
Scatterplot(color divided by Demo)

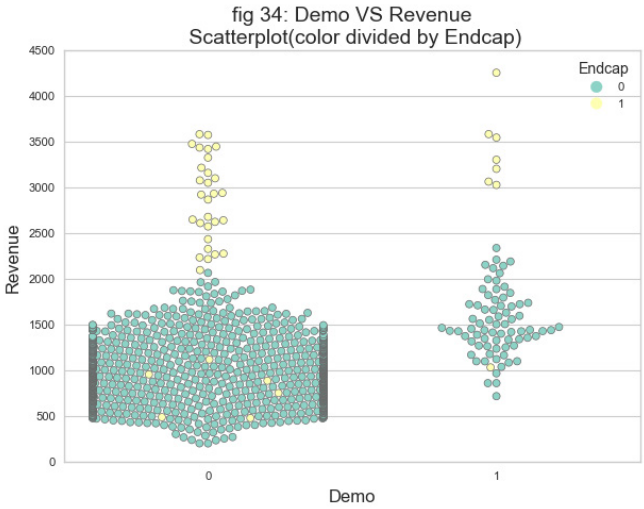


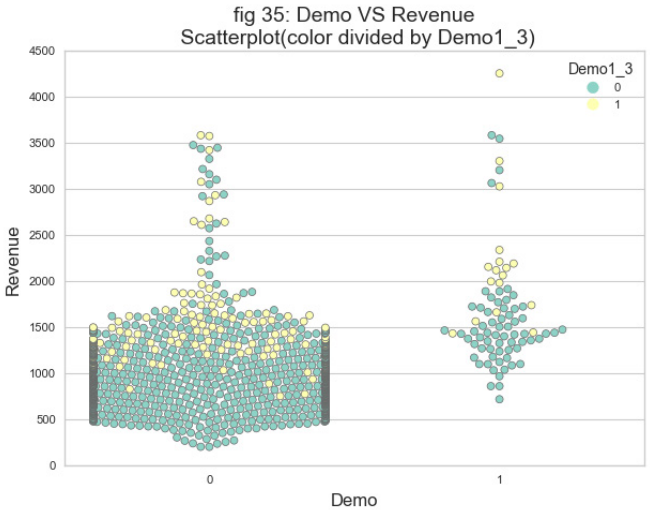


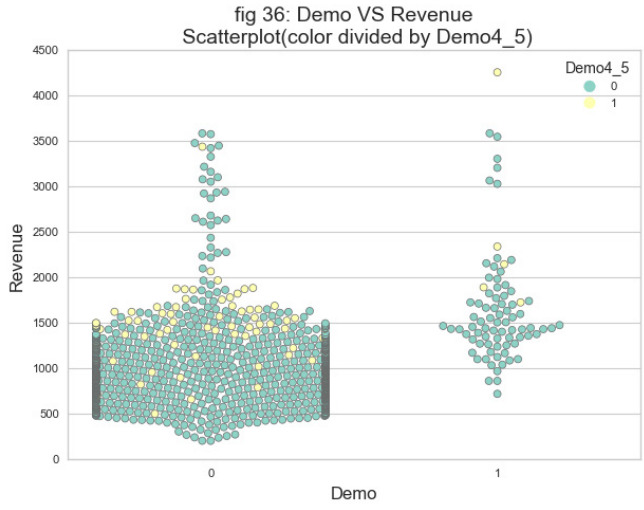


6.1.1.5 Swarm plots of demo divided into groups based on revenue









6.1.1.6 Swarm plots of demo1_3 divided into groups based on revenue

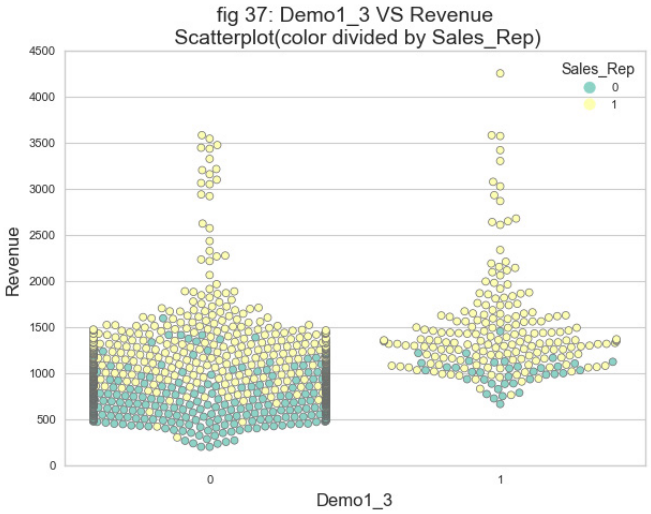


fig 37: Demo1_3 VS Revenue
Scatterplot(color divided by Sales_Rep)

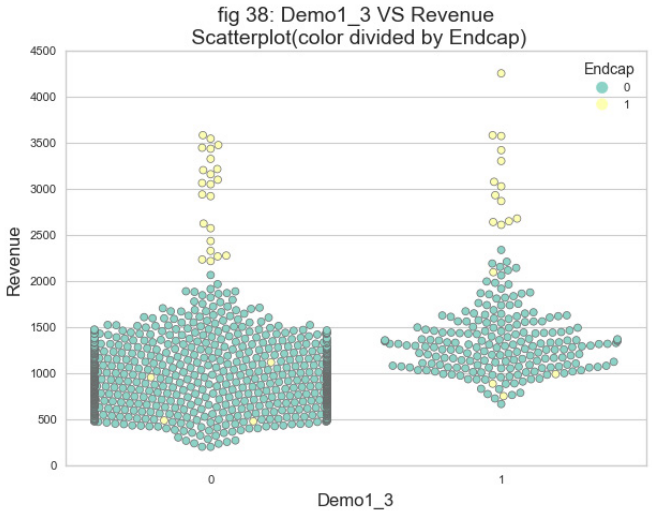


fig 38: Demo1_3 VS Revenue
Scatterplot(color divided by Endcap)

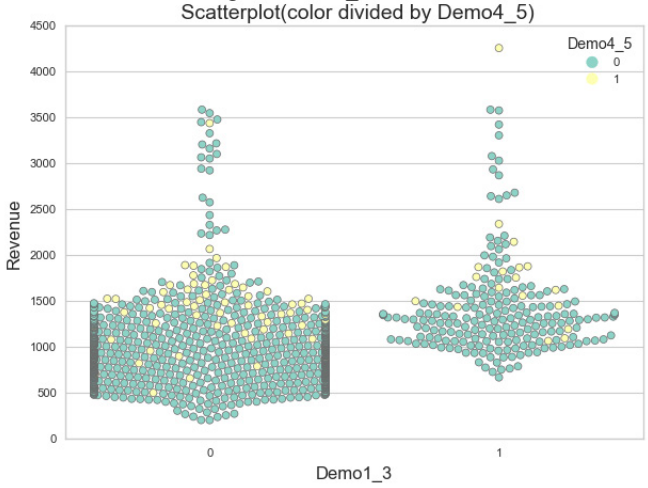


fig 39: Demo1_3 VS Revenue
Scatterplot(color divided by Demo4_5)

6.1.1.7 Swarm plots of demo4_5 divided into groups based on revenue

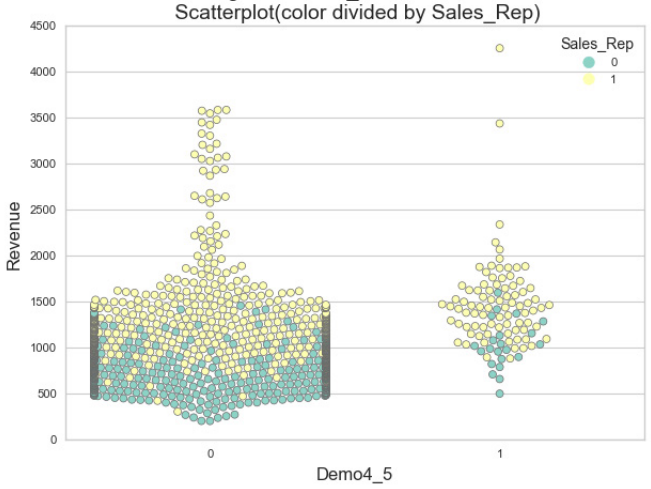


fig 40: Demo4_5 VS Revenue
Scatterplot(color divided by Sales_Rep)

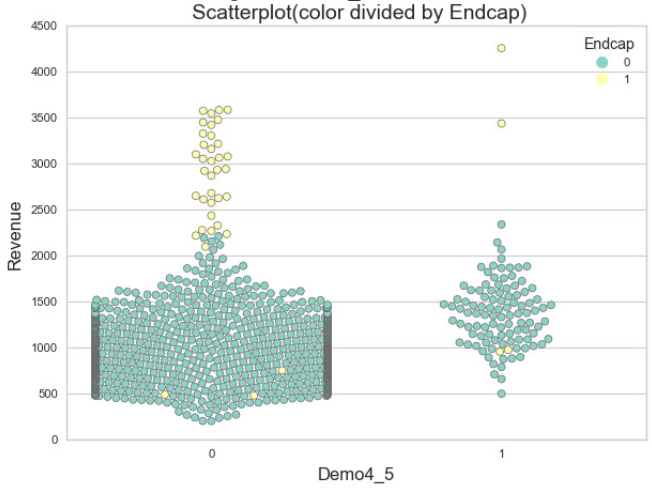
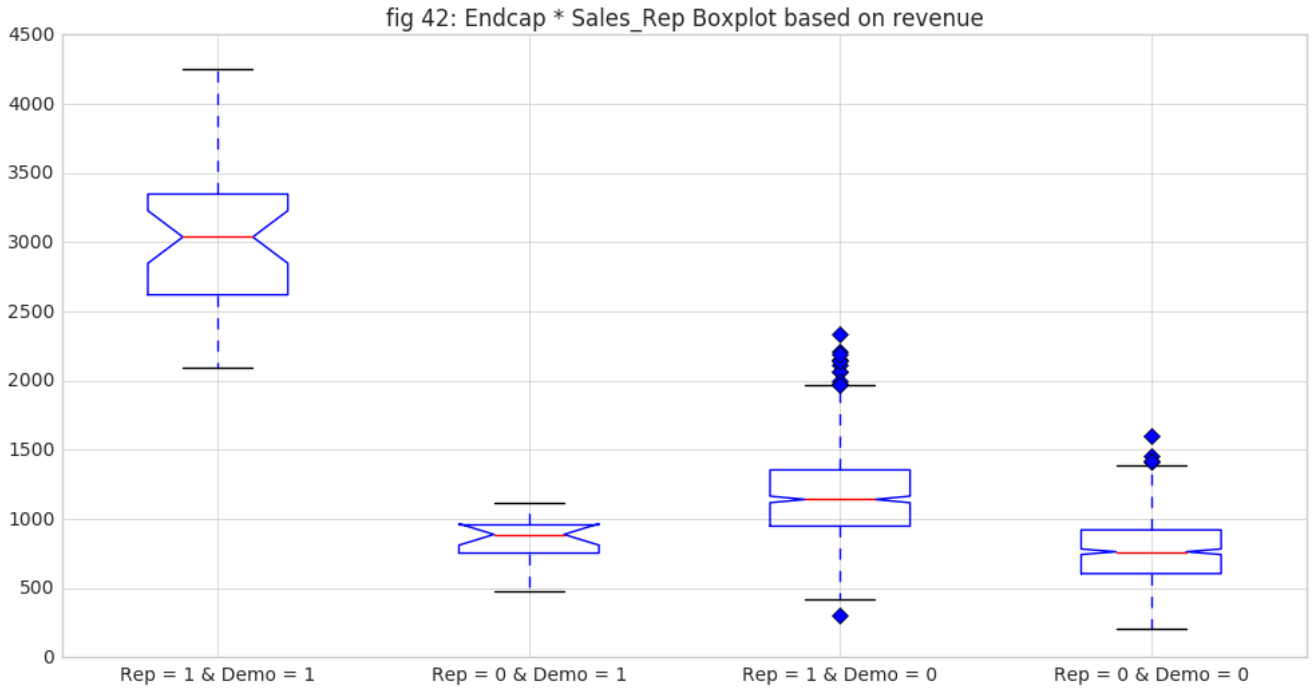


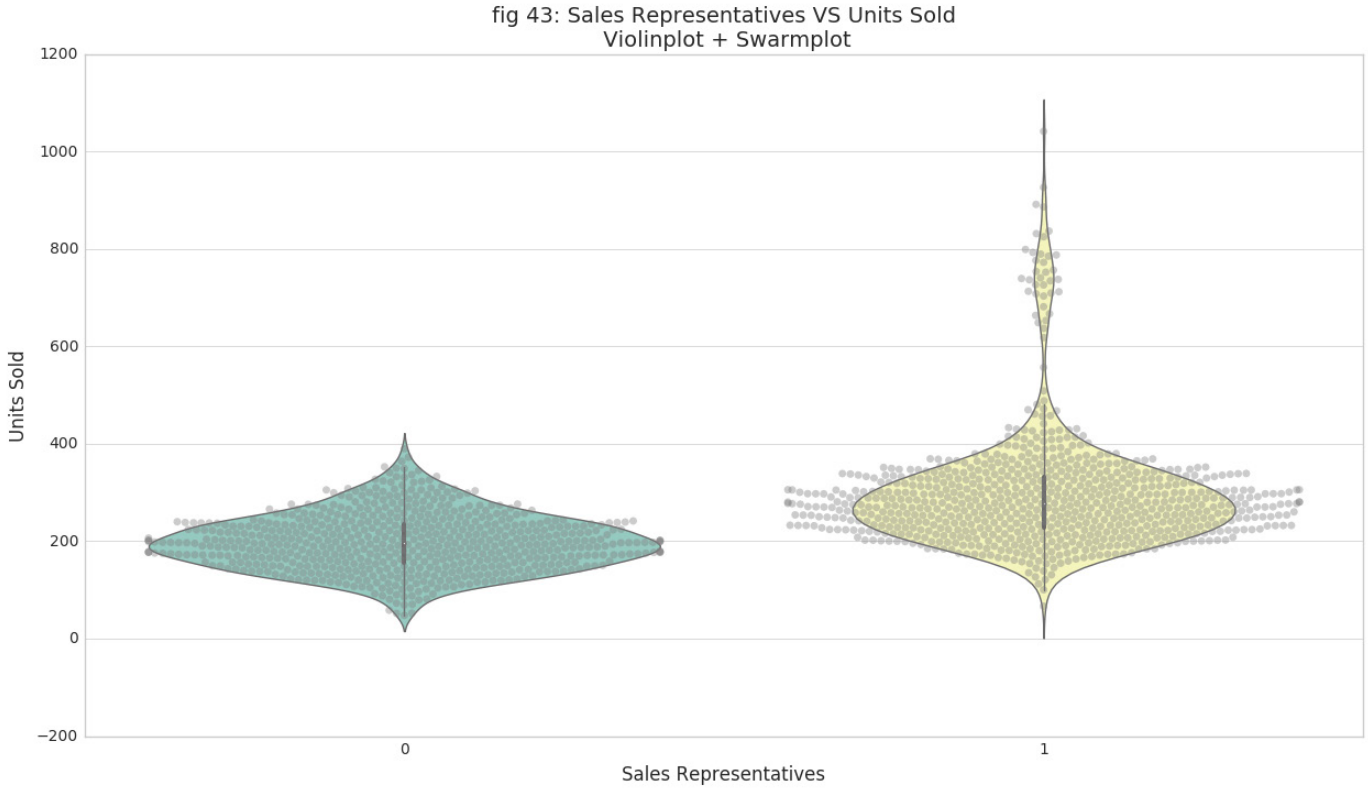
fig 41: Demo4_5 VS Revenue
Scatterplot(color divided by Endcap)

6.1.1.8 Box plots of endcap x sales representatives based on revenue

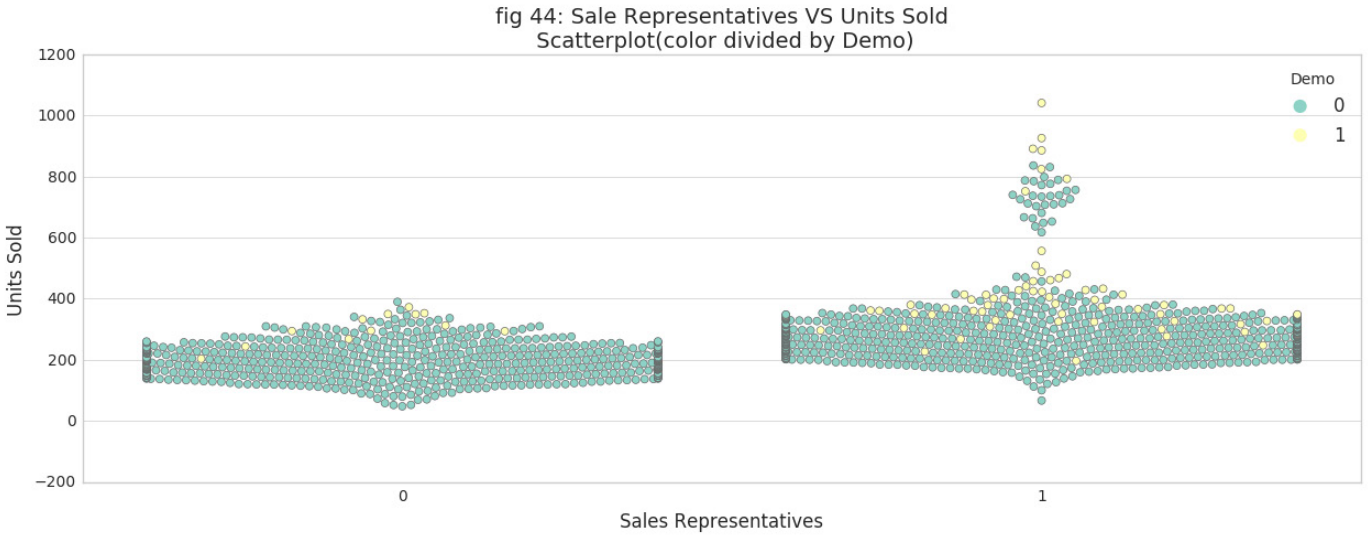


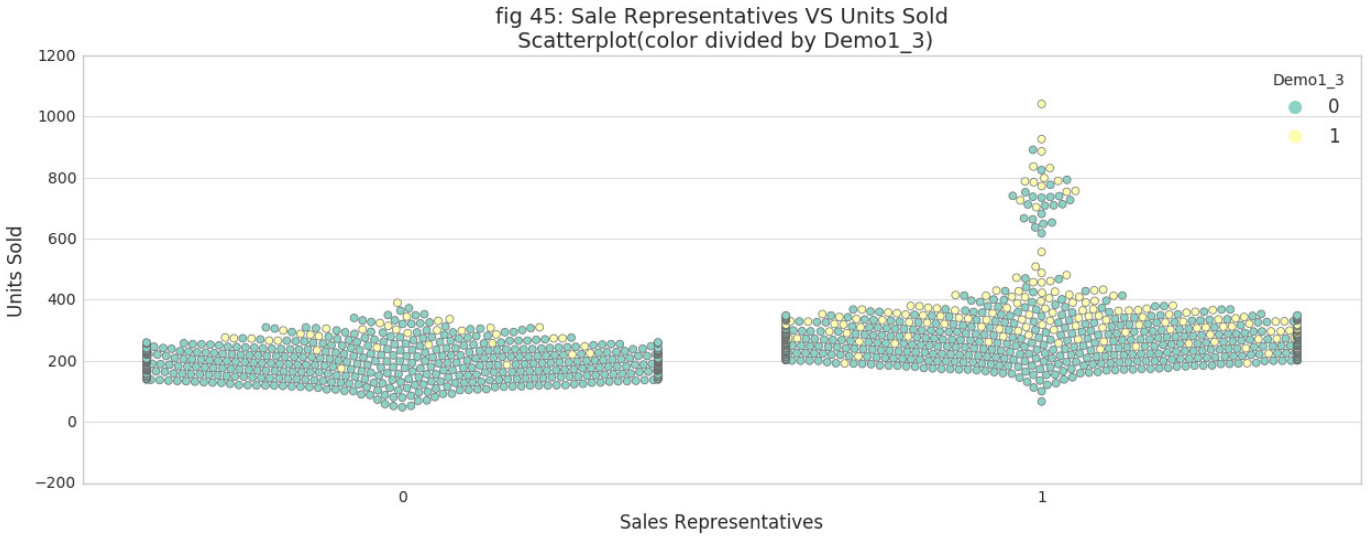
6.1.2 Sales units as dependent variable:

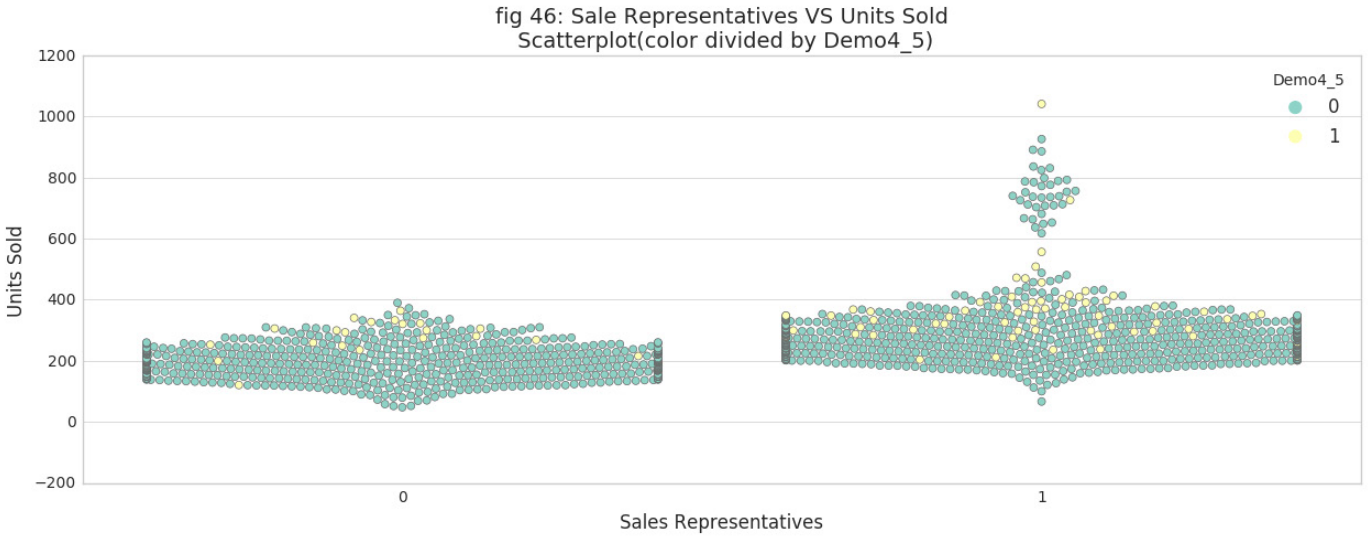
6.1.2.1 Violin plot and swarmplot of sales representatives based on units sold



6.1.2.2 Swarmplot of sales representatives divided into groups based on units sold







6.1.2.3 Swarmplot of endcap divided into groups based on units sold

fig 47: Endcap VS Units Sold
Scatterplot(color divided by Demo)

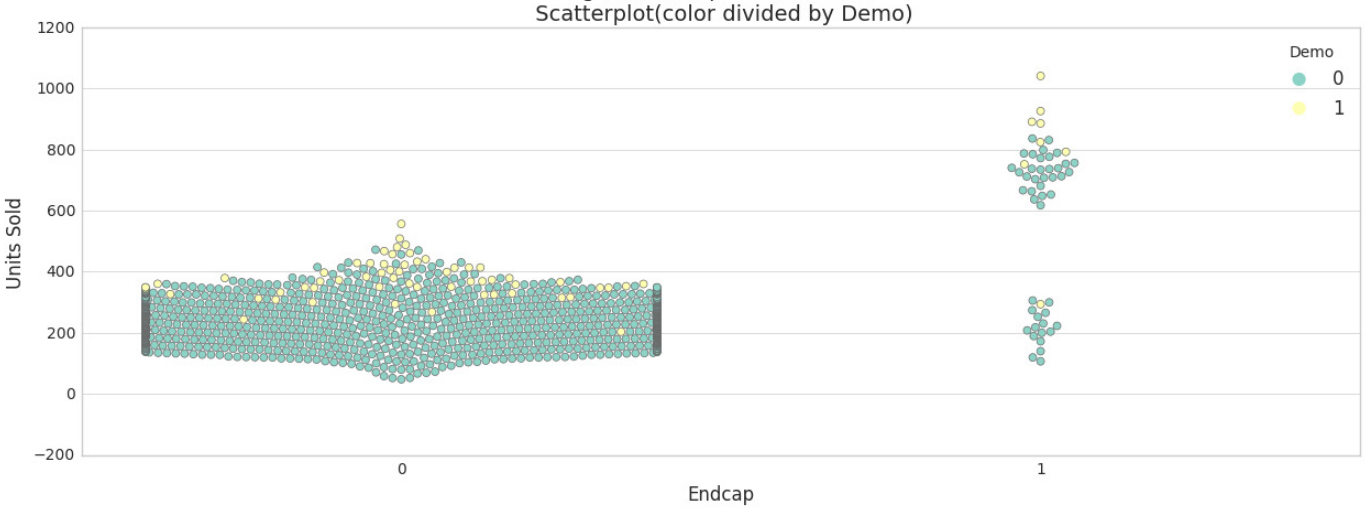


fig 48: Endcap VS Units Sold
Scatterplot(color divided by Demo1_3)

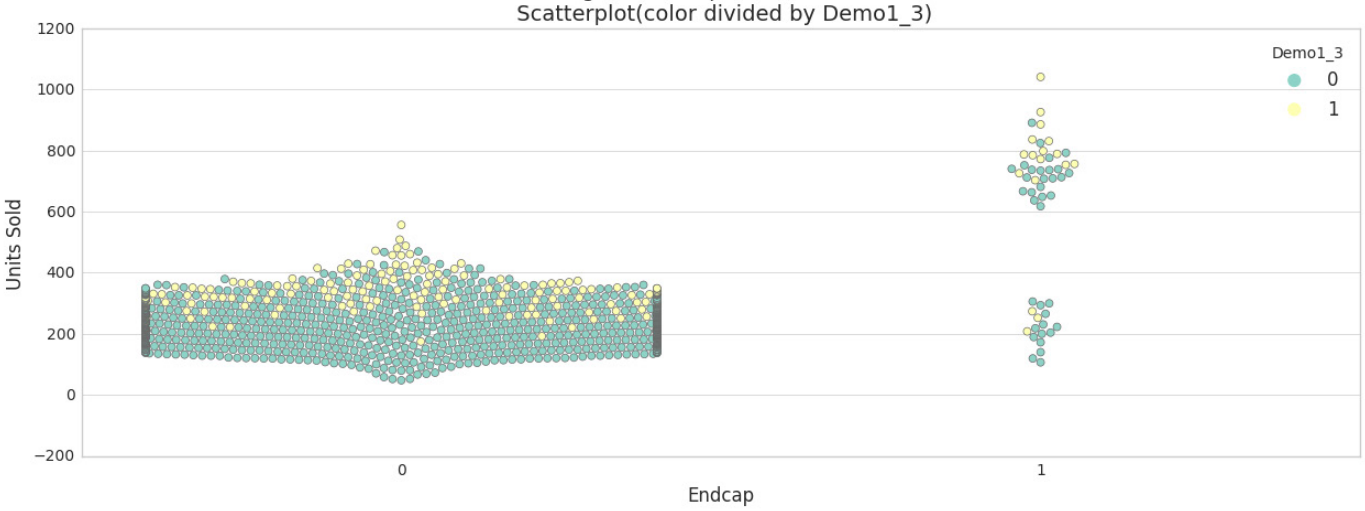
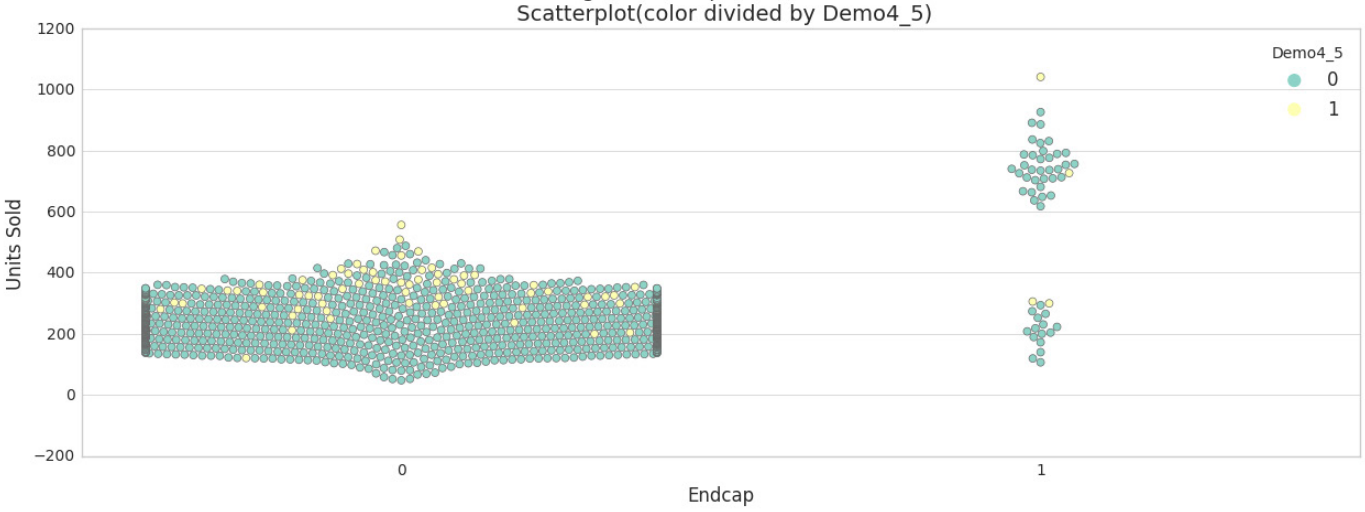
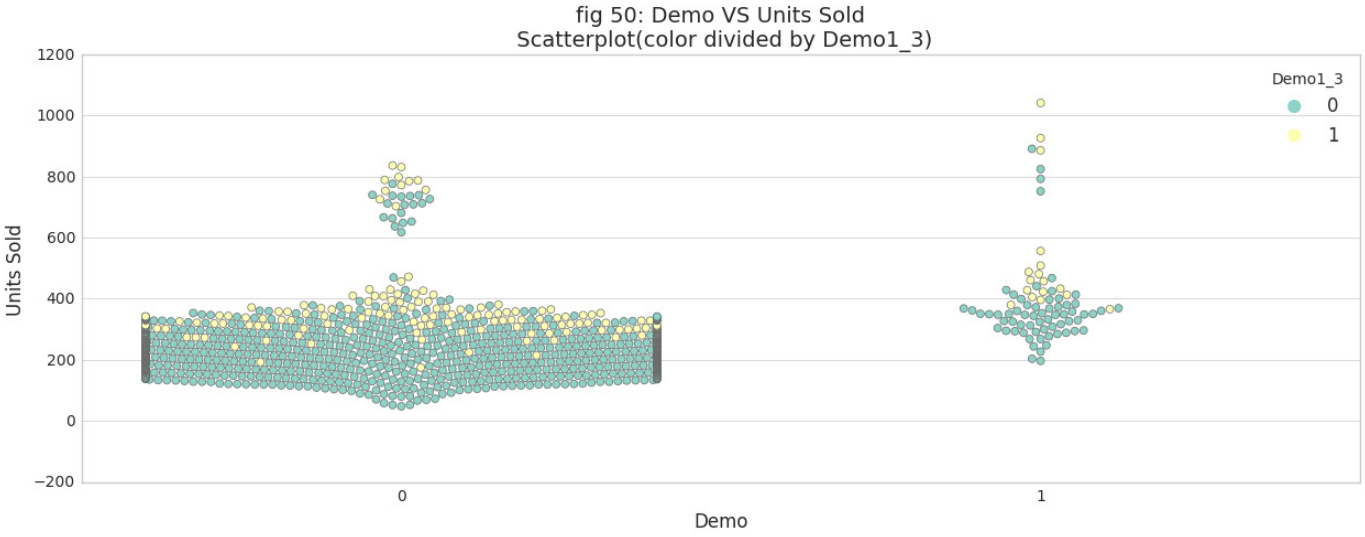
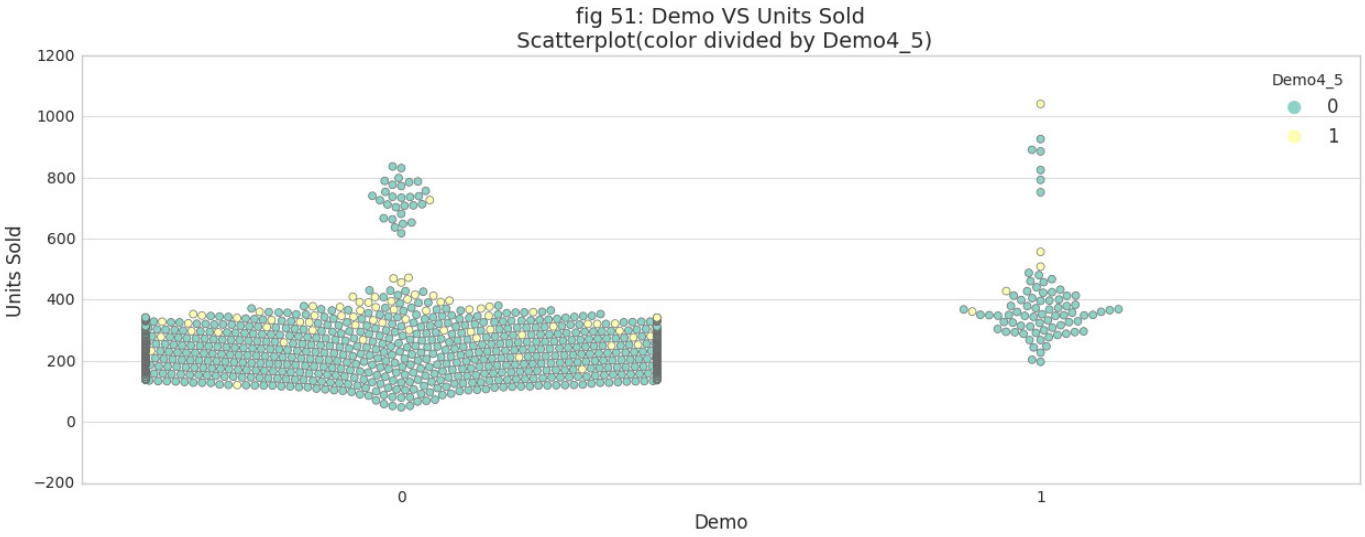


fig 49: Endcap VS Units Sold
Scatterplot(color divided by Demo4_5)

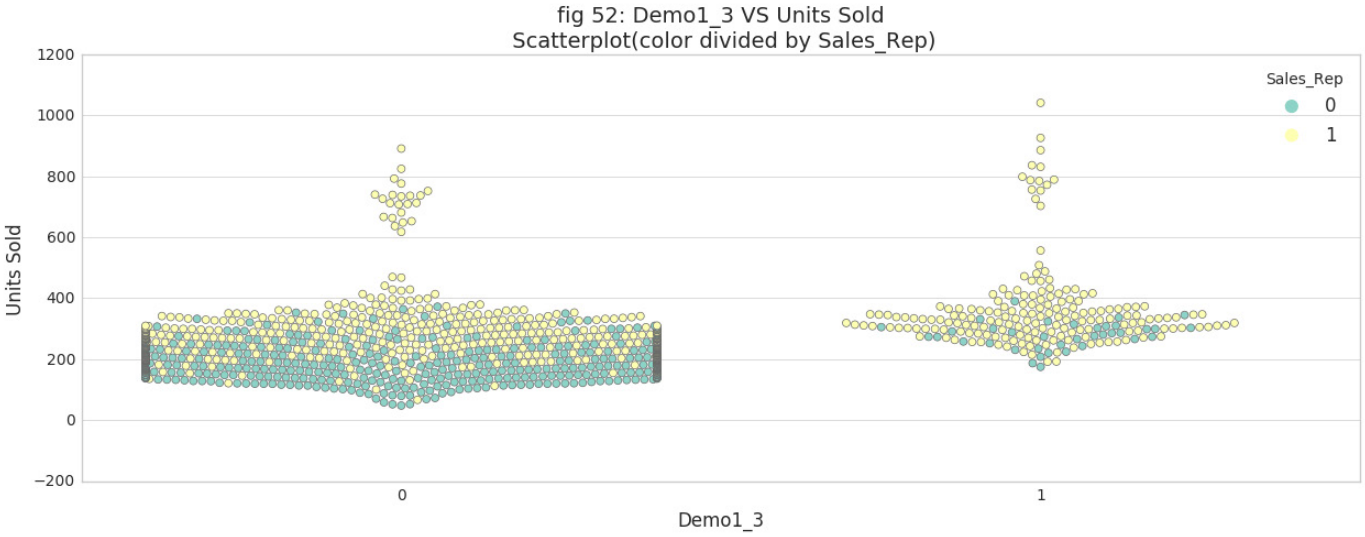


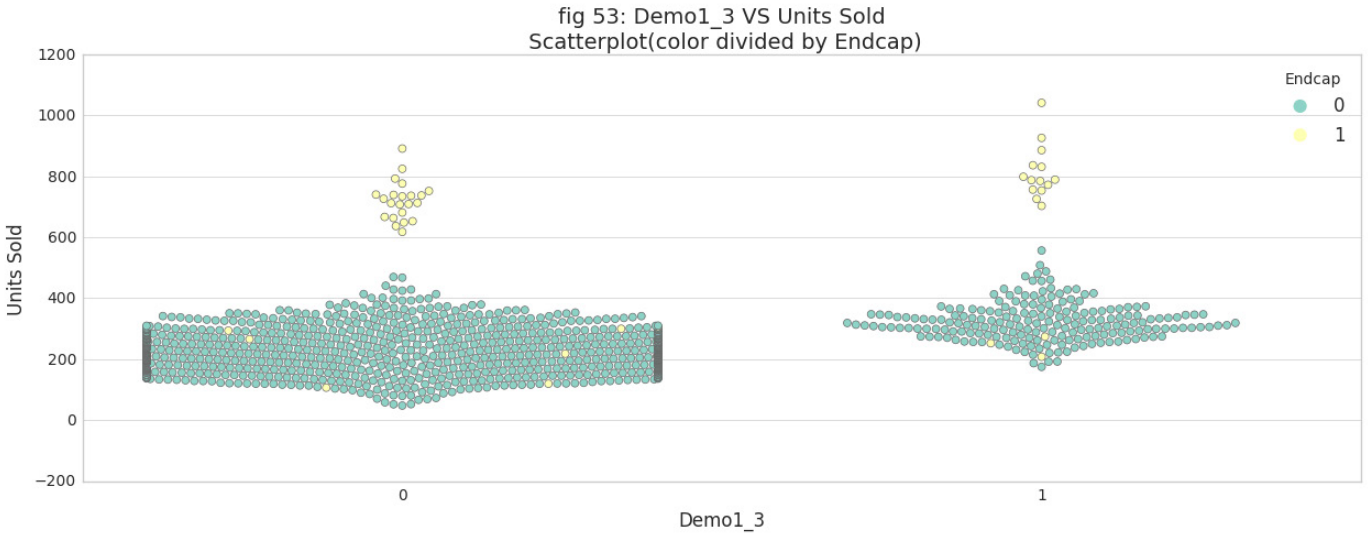
6.1.2.4 Swarmplot of demo divided into groups based on units sold



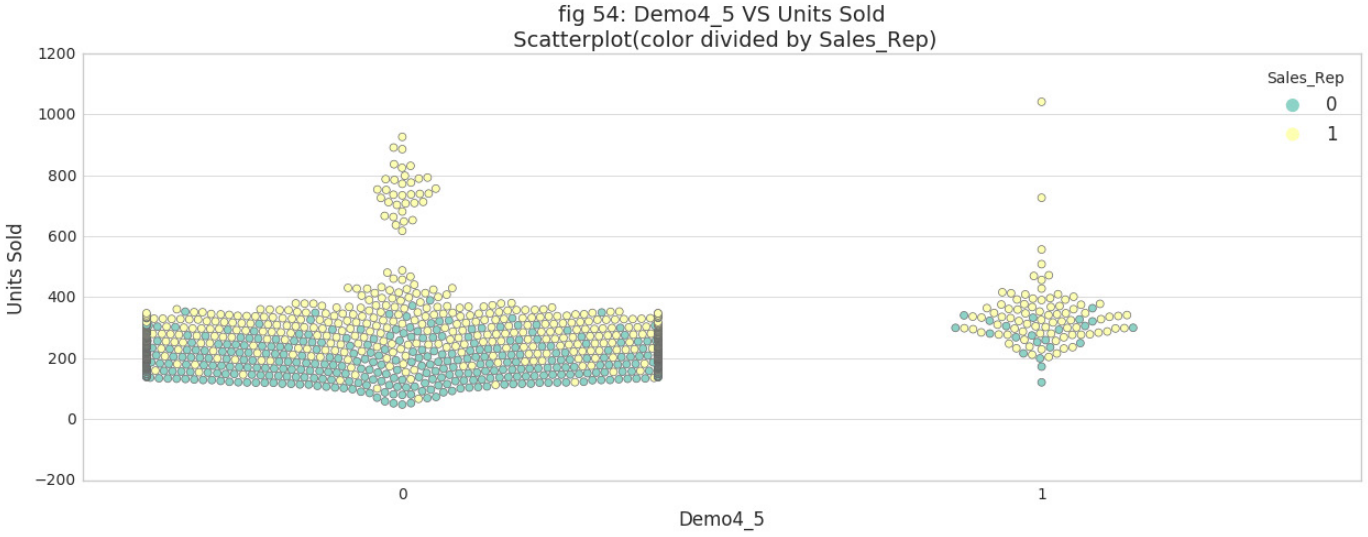


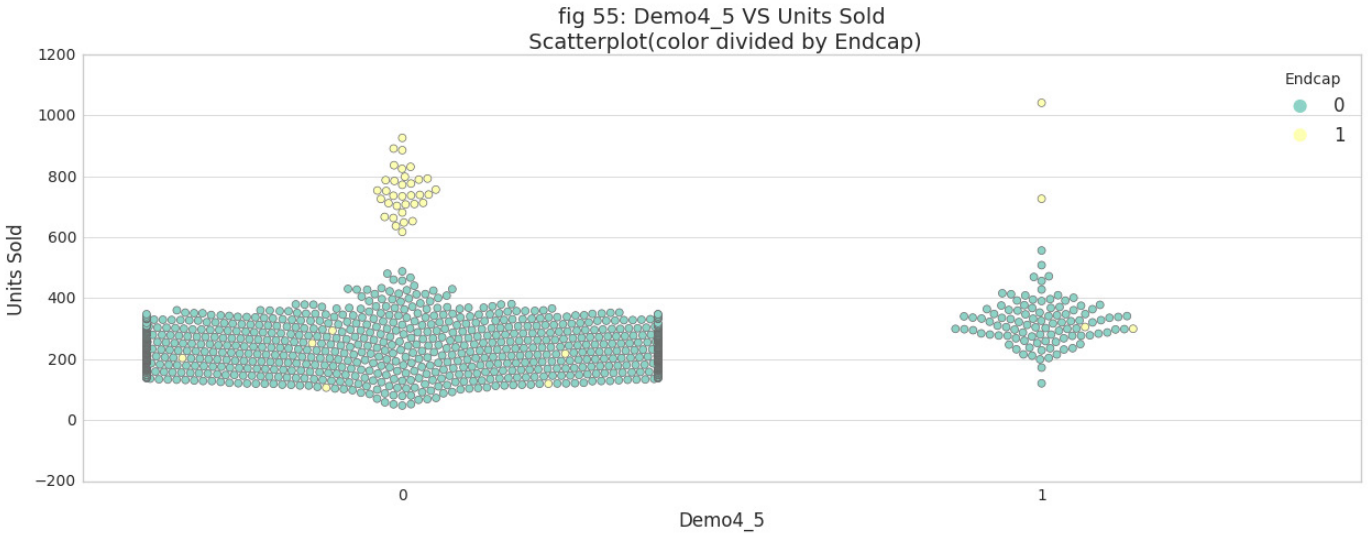
6.1.2.5 Swarmplot of demo1_3 divided into groups based on units sold





6.1.2.6 Swarmplot of demo1_3 divided into groups based on units sold





6.2 Model performance

OLS Regression Results

Dep. Variable:	Unit	R-squared:	0.806
Model:	OLS	Adj. R-squared:	0.805
Method:	Least Squares	F-statistic:	817.1
Date:	Thu, 22 Mar 2018	Prob (F-statistic):	0.00
Time:	12:40:32	Log-Likelihood:	-7357.6
No. Observations:	1386	AIC:	1.473e+04
Df Residuals:	1378	BIC:	1.477e+04
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	276.5735	12.169	22.728	0.000	252.703	300.444
Rep	59.4618	3.011	19.747	0.000	53.555	65.369
Endcap	0.6204	12.077	0.051	0.959	-23.071	24.311
Rep:Endcap	453.8033	14.737	30.793	0.000	424.894	482.713
Price	-22.0664	3.045	-7.248	0.000	-28.039	-16.094
Demo	106.7527	5.700	18.728	0.000	95.571	117.935
Demo1_3	73.3698	3.766	19.482	0.000	65.982	80.757
Demo4_5	74.5520	5.040	14.793	0.000	64.666	84.438

Omnibus:	0.235	Durbin-Watson:	2.092
Prob(Omnibus):	0.889	Jarque-Bera (JB):	0.159
Skew:	-0.010	Prob(JB):	0.923
Kurtosis:	3.048	Cond. No.	59.4

6.3 Anova table of final model

	sum_sq	df	F	PR(>F)
Rep	1.627912e+06	1.0	677.128794	9.317340e-122
Endcap	4.641594e+06	1.0	1930.667985	2.242068e-264
Rep:Endcap	2.279630e+06	1.0	948.210549	7.049985e-159
Price	1.262856e+05	1.0	52.528429	7.039120e-13
Demo	8.432156e+05	1.0	350.734937	6.686836e-70
Demo1_3	9.125070e+05	1.0	379.556671	7.312994e-75
Demo4_5	5.260949e+05	1.0	218.828825	4.566172e-46
Residual	3.312904e+06	1378.0	NaN	NaN