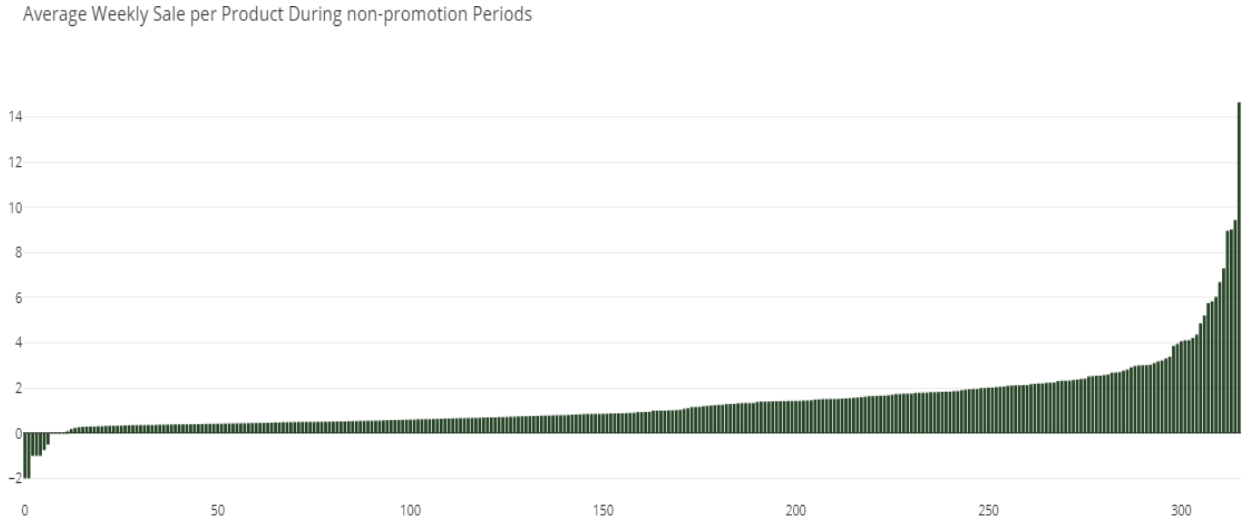
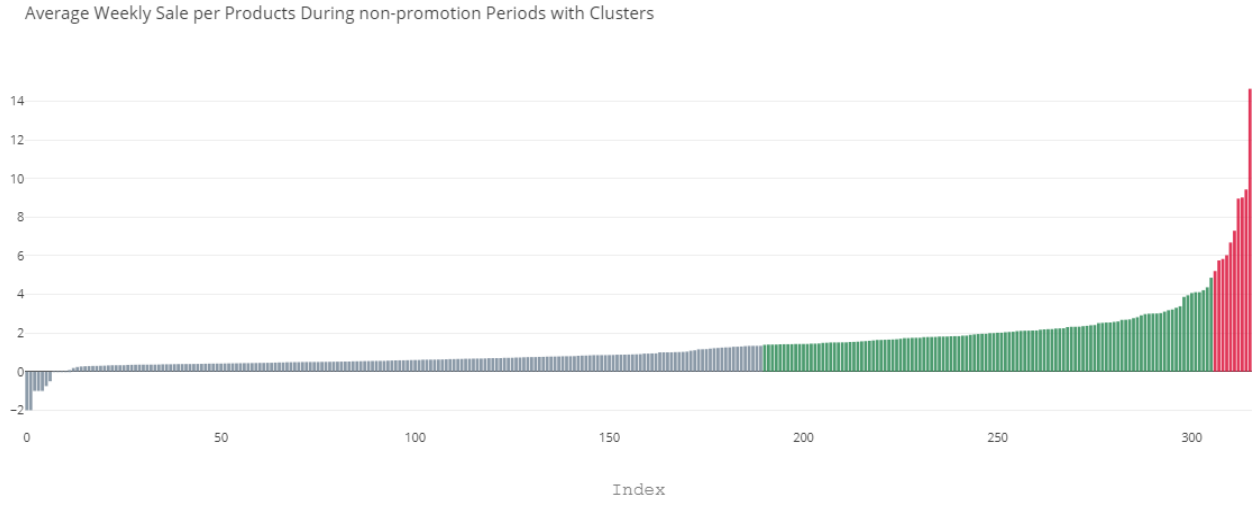
**Anil Ozdemir - Case**

In this case, I will apply a several methods to come up with insights based on the data you supplied to me and going to decide whether the promotions are beneficial while finding the list of questions you are expected to me to find answers. In general, the idea I'm going through is to prove statistically whether promotions work. I used linear regression to do this.

**What are my criteria for separating Fast, Medium and Slow items? Why?**

I use k-means clustering algorithm to cluster these items in terms of their weekly average sales. Sometimes, k-means algorithm does not work well on 1-Dimensional cases (just like our case). However, I cross checked this by looking at the threshold values and calculated that the clusters' minimum, maximum and mean values.



**What are my criteria for separating Fast, Medium and Slow Stores? Why?**

I applied same algorithm just as I separate the products. While doing this, the K-means algorithm identifies k number of centroids (3 in our case) and then allocates every data point to the closest cluster. You can see the graphs in the notebook

**Which items experienced the biggest sale increase during promotions?**

For doing this, i divide the data three-part, weekly sales numbers from fast products, medium and slow respectively. Summary from the lineer regression demonstrates that promotions have significant impact on both of them. We can derive this statement from the look at the p values in the summary ( P> |T|). The p-value for each term tests the null hypothesis that the coefficient is equal to zero (Which means there is no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the dependent variable (Weekly Sales Numbers). In both case “slow, medium and fast products” are affected by the p value, which is near zero. However, coefficients are different, Coefficient of promotions are lower in the slow products which means one unit of change in the predictor variable are lower in slow products while holding other constant. This can be mean lower products are much more positively affected by the promotions compare with faster items. Look at the summaries below: left one is slow products and right one is for fasts products.

```
X = data[data['slow_product']==1]['season']
y = data[data['slow_product']==1]['SalesQuantity']

X2 = sm.add_constant(X)
est = sm.OLS(y, X2)
est2 = est.fit()
print(est2.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          SalesQuantity   R-squared:                       0.005
Model:                            OLS   Adj. R-squared:                  0.005
Method:                 Least Squares   F-statistic:                     264.6
Date:                Wed, 17 Jul 2019   Prob (F-statistic):           2.32e-59
Time:                        23:38:43   Log-Likelihood:            -2.7793e+05
No. Observations:               57735   AIC:                         5.559e+05
Df Residuals:                   57733   BIC:                         5.559e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         20.9508      0.135    154.780      0.000      20.686      21.216
season         5.5106      0.339     16.266      0.000       4.847       6.175
==============================================================================
Omnibus:                    53683.191   Durbin-Watson:                   1.445
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          7348186.425
Skew:                           4.056   Prob(JB):                         0.00
Kurtosis:                      57.670   Cond. No.                         2.81
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly spe
```

```
X = data[data['fast_product']==1]['season']
y = data[data['fast_product']==1]['SalesQuantity']

X2 = sm.add_constant(X)
est = sm.OLS(y, X2)
est2 = est.fit()
print(est2.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          SalesQuantity   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     470.0
Date:                Wed, 17 Jul 2019   Prob (F-statistic):           3.56e-104
Time:                        23:38:33   Log-Likelihood:            -9.4874e+05
No. Observations:              489646   AIC:                         1.897e+06
Df Residuals:                  489644   BIC:                         1.897e+06
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.1894      0.003    453.733      0.000       1.184       1.194
season         0.1415      0.007     21.680      0.000       0.129       0.154
==============================================================================
Omnibus:                   477635.164   Durbin-Watson:                   1.732
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         54683786.270
Skew:                           4.492   Prob(JB):                         0.00
Kurtosis:                      53.987   Cond. No.                         2.80
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly s
```

**Are there stores that have higher promotion reaction?**

Answer of question is yes, look at the summaries below, fast and slow stores are both affected significantly by the promotions while fast store has slightly more coefficient.

```
X = data[data['fast_store']==1]['season']
y = data[data['fast_store']==1]['SalesQuantity']

X2 = sm.add_constant(X)
est = sm.OLS(y, X2)
est2 = est.fit()
print(est2.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          SalesQuantity   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                  0.002
Method:                 Least Squares   F-statistic:                     153.9
Date:                Wed, 17 Jul 2019   Prob (F-statistic):           2.59e-35
Time:                        23:37:54   Log-Likelihood:            -4.2745e+05
No. Observations:               97802   AIC:                         8.549e+05
Df Residuals:                   97800   BIC:                         8.549e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          7.2289      0.067    108.499      0.000       7.098       7.359
season         2.0895      0.168     12.406      0.000       1.759       2.420
==============================================================================
Omnibus:                   147159.945   Durbin-Watson:                   1.321
Prob(Omnibus):                  0.000   Jarque-Bera (JB):        180485810.868
Skew:                           8.955   Prob(JB):                         0.00
Kurtosis:                     212.689   Cond. No.                         2.83
==============================================================================

Warnings:
```

```
X = data[data['slow_store']==1]['season']
y = data[data['slow_store']==1]['SalesQuantity']

X2 = sm.add_constant(X)
est = sm.OLS(y, X2)
est2 = est.fit()
print(est2.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          SalesQuantity   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                  0.002
Method:                 Least Squares   F-statistic:                     523.1
Date:                Wed, 17 Jul 2019   Prob (F-statistic):          1.16e-115
Time:                        23:38:21   Log-Likelihood:            -1.1064e+06
No. Observations:              289266   AIC:                         2.213e+06
Df Residuals:                  289264   BIC:                         2.213e+06
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          4.6449      0.023    206.398      0.000       4.601       4.689
season         1.2849      0.056     22.870      0.000       1.175       1.395
==============================================================================
Omnibus:                   343723.957   Durbin-Watson:                   1.491
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         52555327.205
Skew:                           6.304   Prob(JB):                         0.00
Kurtosis:                      67.819   Cond. No.                         2.81
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly s
```

**Is there any significant difference between promotion impacts of the Fast versus Slow items?**

Not directly apply hypothesis test but coefficients from the slow items are much higher than fast items and both of them affected by the promotions for sure (look at p values). However, coefficient of determination is very low on both case.

**Is there any significant difference between promotion impacts of the Fast versus Slow stores?**

Not directly apply hypothesis test but coefficients from the slow items are very closer to fast items and both of them affected by the promotions for sure (look at p values). However, coefficient of determination is very low on both case either.

**What measure would you use for goodness of fit?**

R-squared = Explained variation / Total variation

R-squared is always between 0 and 100%: and in my cases it is about 0.005 in best case. This means that the model explains little of the variability of the response data around its mean.

**How good is your model developed in step 1?**

I develop not only one model, I develop four model,

First model -> DV: weekly sales, IV : promotion

First model -> DV: weekly sales, IV : promotion, product type

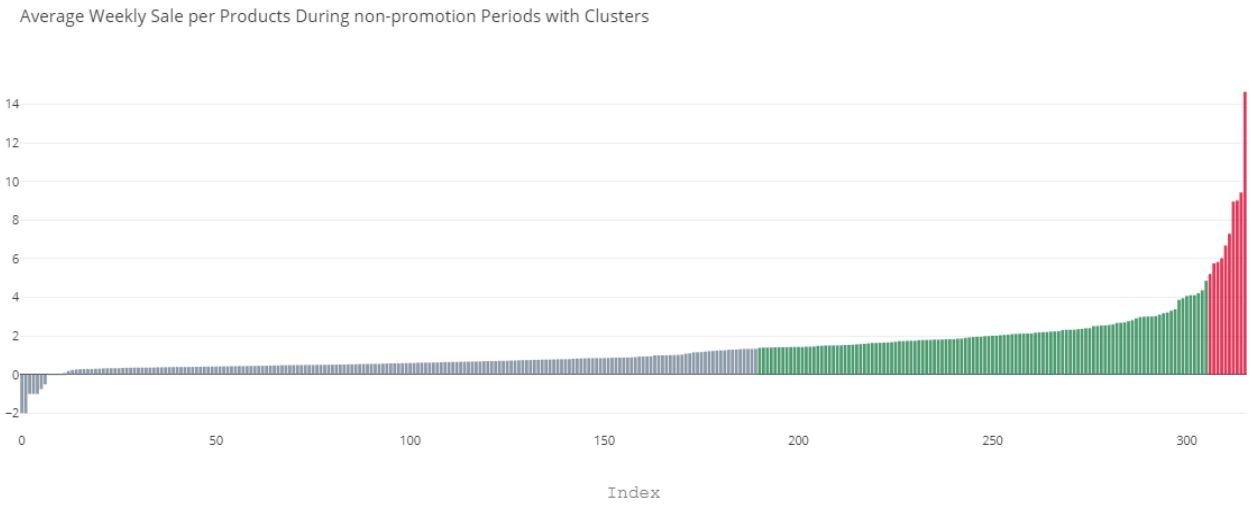
First model -> DV: weekly sales, IV : promotion, store type

First model -> DV: weekly sales, IV : promotion, store type and product type

However, p value from product type and store type greater than 0.05. This means that these type of configurations does not affect the sales and must be remove from the linear equation. Because of that I've select the first model since error is much lower than the other models.

**What are the main problem points causing bad fits?**

Mean square error is 64.3. in my model. This is due to fast items corresponds the very small proportion compare with the small and medium type of products. Therefore, linear model cannot predict that.
When we have an uneven distribution, such as 10 fast product, 90 medium product and 180 small product. This cause bad fits, because linear model can predict small products well but not quite well in other types etc.



**What would you change in step 1?**

I would change my clusterin technique and can be divide data according to the sample size for example 25% for small items, %50 for medium and %25 for fast items.

**Conclusion and Recommendation**

Results from my model demonstrate that promotions have absolute effect on the weekly sales of products. However, type of product nor type of store has not any effect on these sales. In addition, seasonality and trends are highly important in Time series analysis. In this case, I did not investigate these properties because of narrow time horizon and afraid of getting model more complex.

**Is there any data set that you would like to use in addition to tables provided for this assignment? What are those data sets? How would you obtain them?**

- ➔ How many people visit these stores every day and how much percentage of these people transform into new customers because highly average sales of these stores do not mean success every time. Also, with the information of customers we can calculate the promotion retentions etc.