BERKAY AKTÜRK (The images in the PDF file may not appear in good quality.)

Promotion Bump Assignment:

Write a code in R,Python, SAS, Matlab, or software of your choice to do the following.

Attached is the sample data for sales of a company from 2015-01-01 to 2015-08-01. We have gathered information about sales of sample items along with promotion schedule during 2015.

Our goal is to measure the effects of promotion on sales and give recommendations to marketing department for future promotion campaigns.

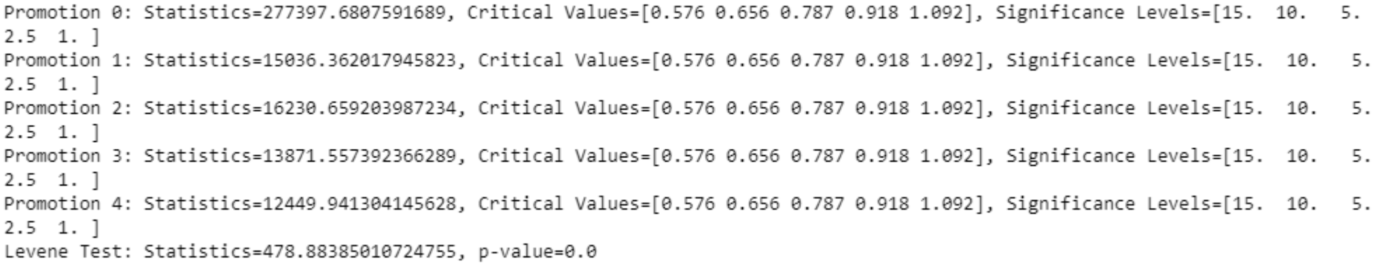
1. Data = Assignment4.1a.csv

The data contains daily sales of sample of items in several stores on a specific time frame.
Negative sale quantities represents returns. Each row represents a sale (or return) activity for an item in a store at a specific day. If a store-item combination has no observation in a certain day you can assume there is no sales for that item at that store at that day.

2. Data=Promotiondates.csv

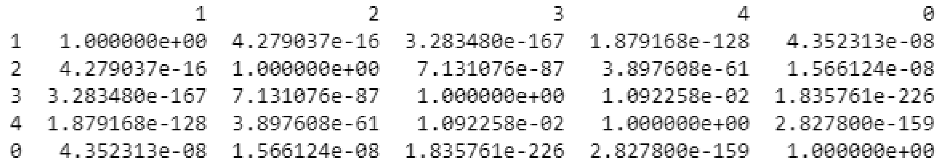
The data contains beginning and the end dates of 6 promotions that took place in 2015.

We can check for statistically significant differences in the changes (averages) in sales quantities corresponding to the relevant promotions and non-promotion periods, similar to the questions mentioned below. ANOVA analysis can be used for this assumption. However, the necessary assumptions must be checked. Since the number of observations is more than 5000, the assumption of normal distribution will be carried out with the Anderson-Darling test.



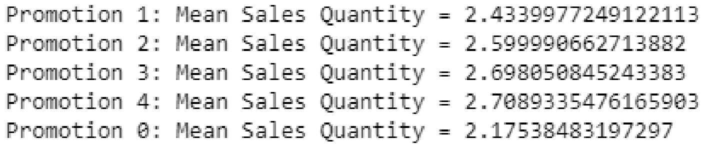
```
Promotion 0: Statistics=277397.6807591689, Critical Values=[0.576 0.656 0.787 0.918 1.092], Significance Levels=[15.  10.   5.
2.5  1. ]
Promotion 1: Statistics=15036.362017945823, Critical Values=[0.576 0.656 0.787 0.918 1.092], Significance Levels=[15.  10.   5.
2.5  1. ]
Promotion 2: Statistics=16230.659203987234, Critical Values=[0.576 0.656 0.787 0.918 1.092], Significance Levels=[15.  10.   5.
2.5  1. ]
Promotion 3: Statistics=13871.557392366289, Critical Values=[0.576 0.656 0.787 0.918 1.092], Significance Levels=[15.  10.   5.
2.5  1. ]
Promotion 4: Statistics=12449.941304145628, Critical Values=[0.576 0.656 0.787 0.918 1.092], Significance Levels=[15.  10.   5.
2.5  1. ]
Levene Test: Statistics=478.88385010724755, p-value=0.0
```

The results of the Anderson-Darling test indicate that the normal distribution is rejected for all promotion groups. Statistically, all values are quite larger than the significance level. This situation shows that the data is not normally distributed, so it may not be appropriate for us to use statistical tests based on normal distribution. In addition, the Levene test shows that variance homogeneity (equal variances) between the groups is rejected (p-value 0.0). This means that the assumptions of classical ANOVA are violated. In such cases, a non-parametric test like the Kruskal-Wallis test, which does not rely on the assumptions of normal distribution and equal variance, can be used. The Kruskal-Wallis Test Result was obtained as Statistics=1771.374455175981, p-value=0.0. There are statistically significant differences between promotions and non-promotion situations. To see which groups differ, Dunn's test can be applied.



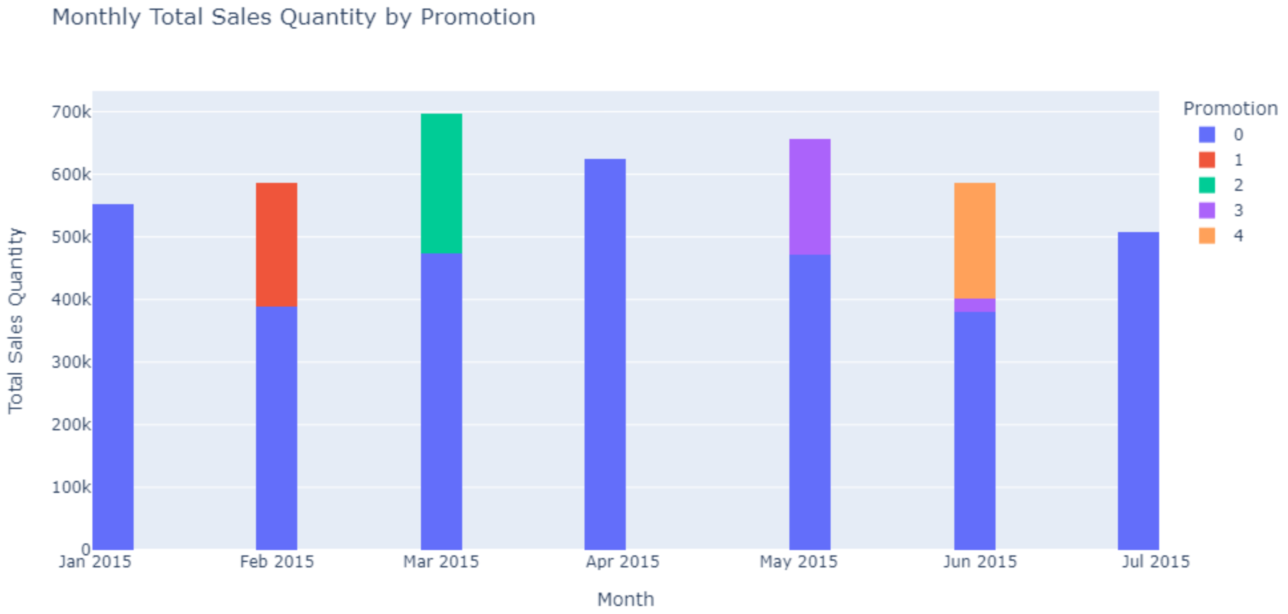
```
                 1             2              3              4              0
1   1.000000e+00  4.279037e-16  3.283480e-167  1.879168e-128   4.352313e-08
2   4.279037e-16  1.000000e+00   7.131076e-87   3.897608e-61   1.566124e-08
3  3.283480e-167  7.131076e-87   1.000000e+00   1.092258e-02  1.835761e-226
4  1.879168e-128  3.897608e-61   1.092258e-02   1.000000e+00  2.827800e-159
0   4.352313e-08  1.566124e-08  1.835761e-226  2.827800e-159   1.000000e+00
```

These results include the outcomes of Dunn's post-hoc test, where each promotion group (1, 2, 3, 4) is compared to the non-promotion group (0). The p-values indicate whether the differences between the groups are statistically significant. In this case, all p-values are quite low, showing that there are significant differences between the groups. The promotion with the highest sales value among the promotions can be seen in the table below. We can say that Promotion 4 has a higher sales quantity compared to other promotions, and the sales quantity decreases when there is no promotion.



```
Promotion 1: Mean Sales Quantity = 2.4339977249122113
Promotion 2: Mean Sales Quantity = 2.599990662713882
Promotion 3: Mean Sales Quantity = 2.698050845243383
Promotion 4: Mean Sales Quantity = 2.7089335476165903
Promotion 0: Mean Sales Quantity = 2.17538483197297
```

Additionally, with the chart below, we can examine the monthly graph of total sales quantities according to the promotion status.

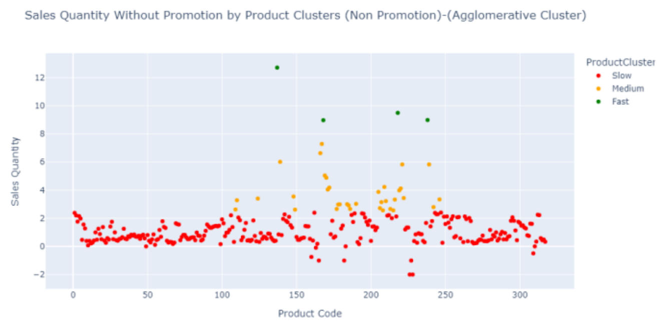


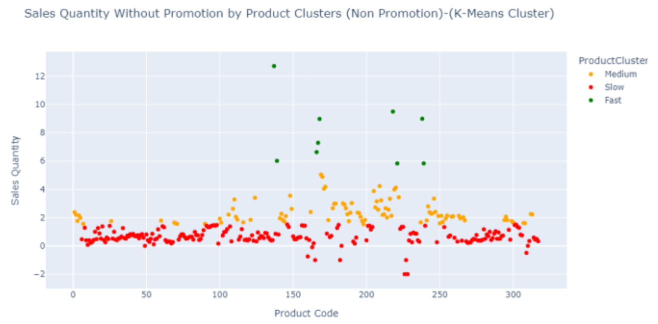
A. Your goal is to model the effect of promotion on products and stores. At this stage only use the data in the file Assignment4.1a.csv and base your model using the first 4 promotions.

In order to answer below questions you should divide product and stores into 3 clusters each. Product with higher average weekly sale per store during non-promotion periods will be called “Fast items” and items with lower weekly average sale per store will be labeled as “Slow items”, items in between will be called “Medium items”. Grouping parameters selection is left to you. Apply similar approach to Stores as well.

a. What are your criteria for separating Fast, Medium and Slow items? Why?

There are many algorithms you can use for clustering. The K-means algorithm and Agglomerative Cluster algorithm are among the most popular clustering algorithms and have thus been used in this example. However, many other clustering algorithms are also available. Among them are Hierarchical Clustering, DBSCAN, Spectral Clustering, Mean-Shift Clustering, OPTICS, and many more. Yet each method tends to produce similar and close results. Moreover, upon examining the weekly average values of product and store observations, it has been decided that the K-Means algorithm is suitable. The criteria for separating "Fast," "Medium," and "Slow" products typically rely on the products' sales rates or sales volumes. Here, the criterion determining the classification of a product as "Fast," "Medium," or "Slow" is the product's average sales quantity during non-promotion weeks. This is because promotion periods generally increase product sales, and therefore may not reflect the normal sales rate. However, comparing these two situations by calculating the weekly average sales during the promotion period would be quite beneficial.

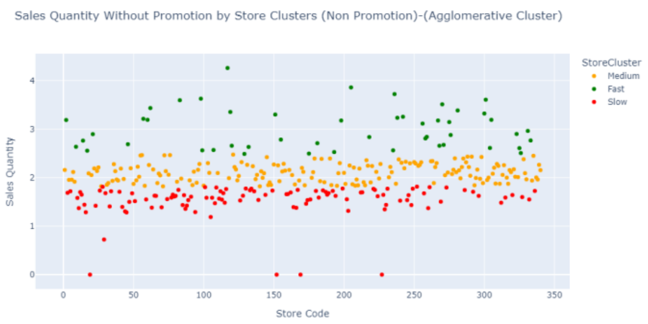


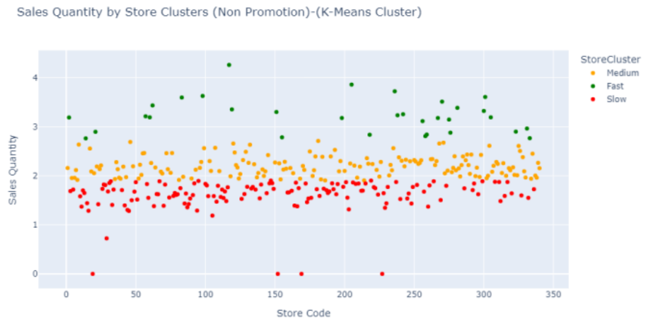


When examining these graphs, we can see that most products are classified as "Slow" during the periods when there are no promotions.

b. What are your criteria for separating Fast, Medium and Slow Stores? Why?

Similarly, the criteria for separating stores into "Fast", "Medium", and "Slow" categories are usually based on the stores' sales volumes or sales rates. Here, the criterion that determines whether a store is classified as "Fast", "Medium", or "Slow" is the store's average weekly sales quantity. This is used as it best reflects the overall performance of each store. Based on these results, the K-Means algorithm will suffice to meet our clustering needs.

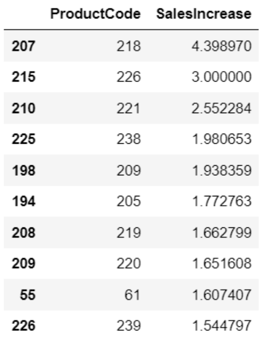




When there is no promotion, it can be seen that the number of slow and high stores is similar compared to the products. This situation varies according to the stores, but there are various stores with high weekly sales averages.

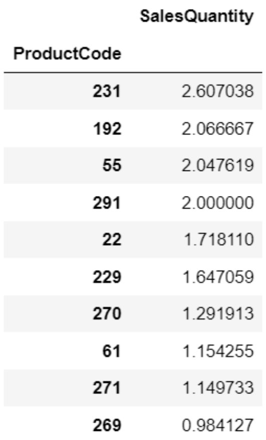
c. Which items experienced the biggest sale increase during promotions?

When the differences in sales quantities are taken before and during promotions, the following table results can be reached, and we can see the ten product codes that experienced the most sales increase during promotions.



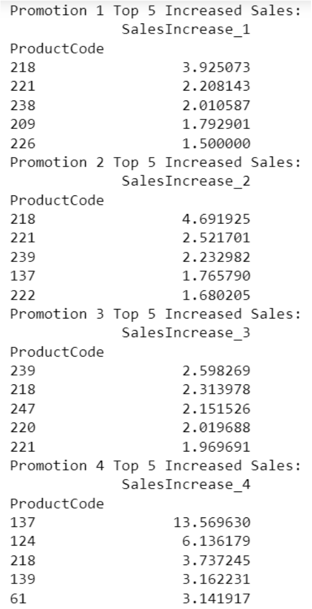
	ProductCode	SalesIncrease
207	218	4.398970
215	226	3.000000
210	221	2.552284
225	238	1.980653
198	209	1.938359
194	205	1.772763
208	219	1.662799
209	220	1.651608
55	61	1.607407
226	239	1.544797

When we take the difference in sales quantities of products during promotions and when there are no promotions, and divide it by the sales quantity of products when there are no promotions, we reach the following ratio table. Here, it can be said that the ratios over 1 are products that experienced high sales during promotions.



	SalesQuantity
ProductCode	
231	2.607038
192	2.066667
55	2.047619
291	2.000000
22	1.718110
229	1.647059
270	1.291913
61	1.154255
271	1.149733
269	0.984127

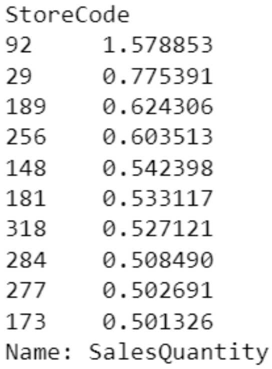
Finally, when each promotion is compared with sales at dates before the promotion was applied, the following table is obtained. By subtracting the number of products sold when there are no promotions from the number of products sold for each promotion, we can examine the products with the highest sales. Those that are common with the products listed in the table above can be referred to as products with high sales.



```
Promotion 1 Top 5 Increased Sales:
             SalesIncrease_1
ProductCode
218                 3.925073
221                 2.208143
238                 2.010587
209                 1.792901
226                 1.500000
Promotion 2 Top 5 Increased Sales:
             SalesIncrease_2
ProductCode
218                 4.691925
221                 2.521701
239                 2.232982
137                 1.765790
222                 1.680205
Promotion 3 Top 5 Increased Sales:
             SalesIncrease_3
ProductCode
239                 2.598269
218                 2.313978
247                 2.151526
220                 2.019688
221                 1.969691
Promotion 4 Top 5 Increased Sales:
             SalesIncrease_4
ProductCode
137                13.569630
124                 6.136179
218                 3.737245
139                 3.162231
61                  3.141917
```

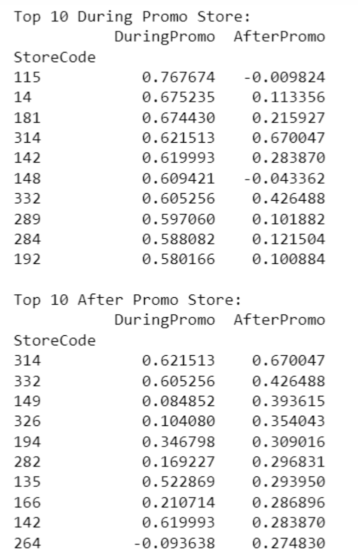
d. Are there stores that have higher promotion reaction?

We can see the top ten stores that have responded most strongly to the promotions below. When these proportionally calculated values are examined, generally Store 92 has been significantly affected by the promotions, showing an increase in sales.



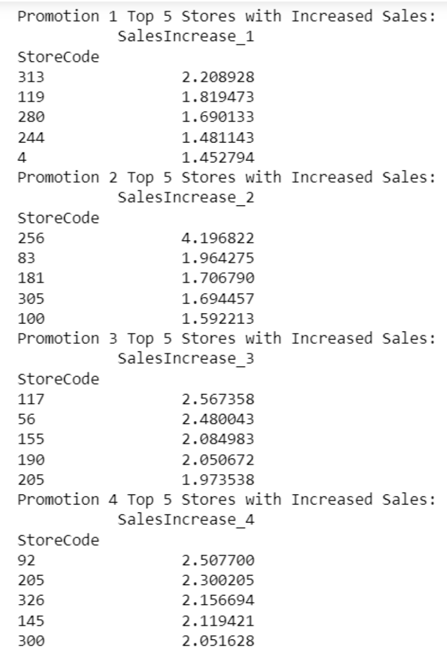
```
StoreCode
92      1.578853
29      0.775391
189     0.624306
256     0.603513
148     0.542398
181     0.533117
318     0.527121
284     0.508490
277     0.502691
173     0.501326
Name: SalesQuantity
```

The comparison of the stores' product sales ratios before, during, and after the promotions can be seen in the tables below. In the first table, we can examine the stores with the highest sales during the promotion and the sales ratios of these stores after the promotion. Except for Store 314, all stores with high sales are showing a decrease in sales after the promotion. In the second table, we can see the top ten stores showing the most sales after the promotion and the ratio of sales values during the promotion. Most of these stores (314,149,326,282,166,264) appear to have increased their sales after the promotion. Among the main reasons for the increase in sales in these stores after the promotion are the widespread introduction of promotions, consumers becoming more informed about the products during the promotional period, and possibly developing brand loyalty. The strategy pursued by these stores, especially Store 314, should be examined.



```
Top 10 During Promo Store:
           DuringPromo  AfterPromo
StoreCode
115           0.767674   -0.009824
14            0.675235    0.113356
181           0.674430    0.215927
314           0.621513    0.670047
142           0.619993    0.283870
148           0.609421   -0.043362
332           0.605256    0.426488
289           0.597060    0.101882
284           0.588082    0.121504
192           0.580166    0.100884

Top 10 After Promo Store:
           DuringPromo  AfterPromo
StoreCode
314           0.621513    0.670047
332           0.605256    0.426488
149           0.084852    0.393615
326           0.104080    0.354043
194           0.346798    0.309016
282           0.169227    0.296831
135           0.522869    0.293950
166           0.210714    0.286896
142           0.619993    0.283870
264          -0.093638    0.274830
```

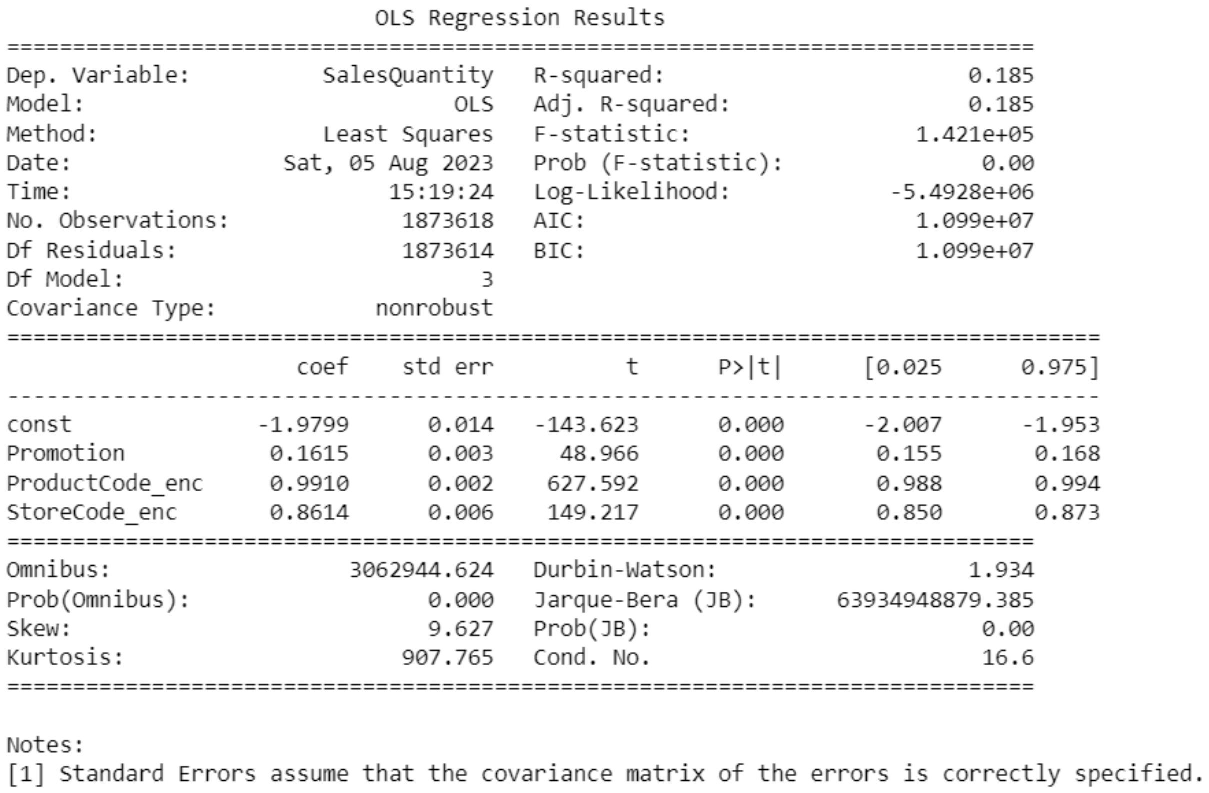
Finally, by taking the difference between the sales values of the stores during the promotion and the sales values before the promotion, we can see the top 5 stores with the highest sales. It includes the sales values of stores for each promotion. Generally, stores show different sales according to the promotion. We can see in the table below that each store has a different sales value according to each promotion, and that stores cannot have similar sales in every promotion.



```
Promotion 1 Top 5 Stores with Increased Sales:
           SalesIncrease_1
StoreCode
313               2.208928
119               1.819473
280               1.690133
244               1.481143
4                 1.452794
Promotion 2 Top 5 Stores with Increased Sales:
           SalesIncrease_2
StoreCode
256               4.196822
83                1.964275
181               1.706790
305               1.694457
100               1.592213
Promotion 3 Top 5 Stores with Increased Sales:
           SalesIncrease_3
StoreCode
117               2.567358
56                2.480043
155               2.084983
190               2.050672
205               1.973538
Promotion 4 Top 5 Stores with Increased Sales:
           SalesIncrease_4
StoreCode
92                2.507700
205               2.300205
326               2.156694
145               2.119421
300               2.051628
```

e. What is the biggest effect explaining sales change during promotions?

The type of promotion, such as discount rate, gifts, loyalty points, can affect sales in different ways. The stores' location and type may influence the reaction to promotions. However, if we want to examine this situation based on the data we have, we can look at the regression table below.



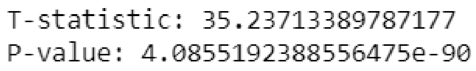
```
                            OLS Regression Results
==============================================================================
Dep. Variable:          SalesQuantity   R-squared:                       0.185
Model:                            OLS   Adj. R-squared:                  0.185
Method:                 Least Squares   F-statistic:                 1.421e+05
Date:                Sat, 05 Aug 2023   Prob (F-statistic):               0.00
Time:                        15:19:24   Log-Likelihood:            -5.4928e+06
No. Observations:             1873618   AIC:                         1.099e+07
Df Residuals:                 1873614   BIC:                         1.099e+07
Df Model:                           3
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
const              -1.9799      0.014   -143.623      0.000      -2.007      -1.953
Promotion           0.1615      0.003     48.966      0.000       0.155       0.168
ProductCode_enc     0.9910      0.002    627.592      0.000       0.988       0.994
StoreCode_enc       0.8614      0.006    149.217      0.000       0.850       0.873
==============================================================================
Omnibus:                  3062944.624   Durbin-Watson:                   1.934
Prob(Omnibus):                  0.000   Jarque-Bera (JB):      63934948879.385
Skew:                           9.627   Prob(JB):                         0.00
Kurtosis:                     907.765   Cond. No.                         16.6
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Looking at the results of this multiple regression model, we can see that promotions, product codes, and store codes have a significant effect on sales quantities. However, the R-squared value of the model is found to be 0.185. This shows that our model explains only 18.5% of the variance in sales quantity. Thus, our model can only explain roughly one-fifth of the change in sales quantity. The coefficient of the promotion variable is 0.1615, showing that promotions have a positive effect on sales quantity. That is, when the promotion situation changes, the sales quantity increases by approximately 0.16 units. The coefficients of the product code and store code variables are also positive, at 0.9910 and 0.8614, respectively. This shows that as the average sales quantity of a particular product code or store code increases, the overall sales quantity will also increase. However, since these coefficients have a high degree of variance, we must interpret these results carefully.

Based on all these results, we can say that promotions, product codes, and store codes have a significant effect on sales quantities. However, we must not forget that our model explains only 18.5% of the variance in sales quantity. This indicates that our model does not consider other potential factors affecting sales quantities. Therefore, we should consider adding other variables to our model to better explain sales quantity. An answer to the question at the bottom has been provided through these variables. Also, since product and store codes are not continuous variables and their increase or decrease does not carry meaning, interpreting the model and acting according to this model will not be correct. I can only answer the question in this way.

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

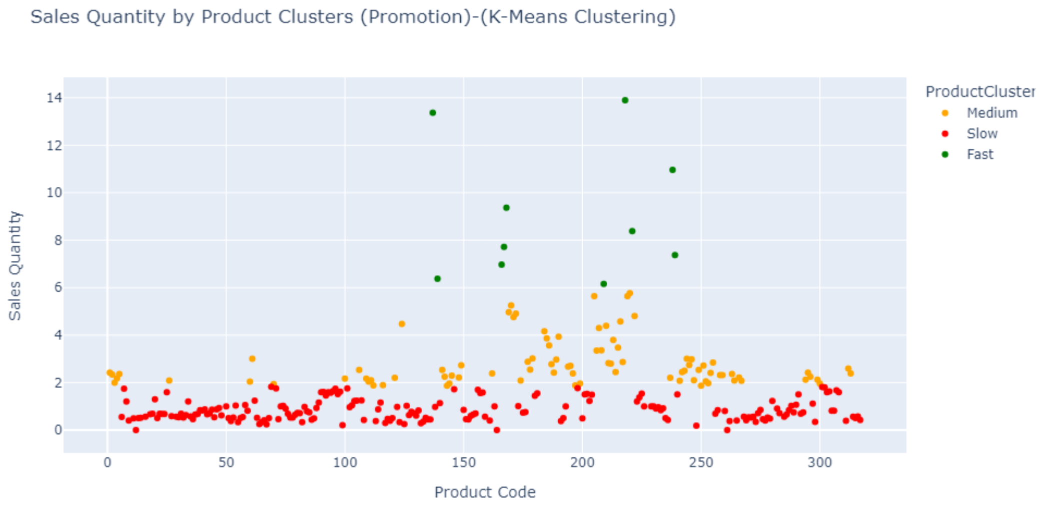
When we look at whether there is a difference between fast and slow products based on the cluster created when there is a promotion or no promotion, it is not possible for the result to be insignificant.



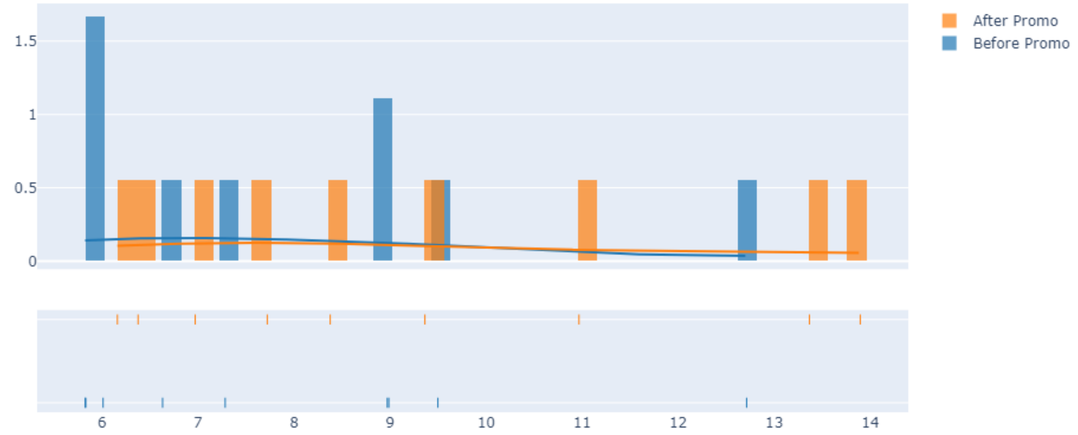
```
T-statistic: 35.23713389787177
P-value: 4.0855192388556475e-90
```

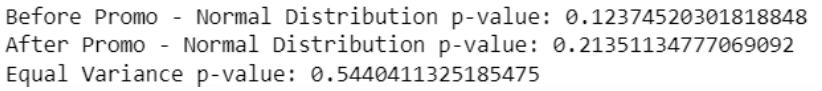
The t-statistic of 35.23713389787177 and a very small P-value (4.0855192388556475e-90, nearly zero) indicate that there is a statistically significant difference between the effects of promotions on "Fast" products versus "Slow" products. In other words, the average effect of promotions on "Fast" items is significantly different compared to "Slow" items. Considering the t-statistic is positive, we can say that promotions have a greater effect on "Fast" items than on "Slow" items.

A better question under these conditions might be; to look at whether there is a statistically significant difference between Fast products when there is a promotion and when there is no promotion. This could lead us to the solution of a more important problem. Below, we can see the cluster of weekly average sales values in the case of a promotion.



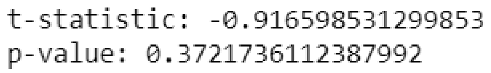
The small number of products with high average weekly sales when there is a promotion and when there is not (Green dots) may lead to an unsuccessful result. Before we test this hypothesis, we should first examine the assumptions.





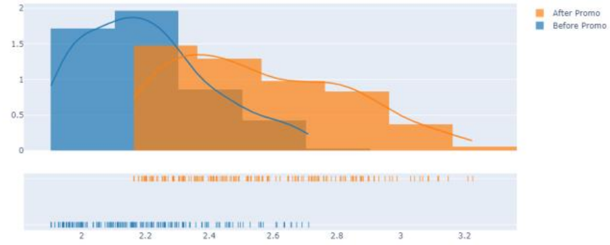
```
Before Promo - Normal Distribution p-value: 0.12374520301818848
After Promo - Normal Distribution p-value: 0.21351134777069092
Equal Variance p-value: 0.5440411325185475
```

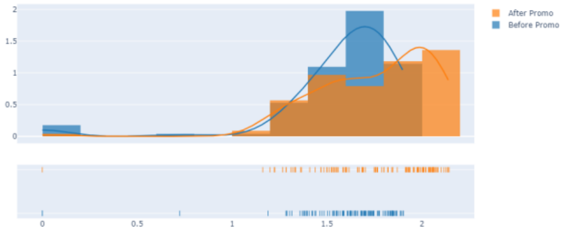
Looking at these results, we can say that the data is normally distributed both when there is no promotion and when there is a promotion (Shapiro-Wilk test p-values > 0.05), and that the variances are equal between both groups (Levene test p-value > 0.05). In this case, a t-test can be used.

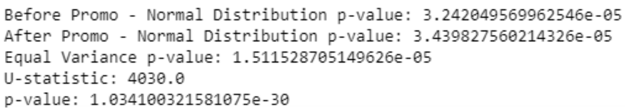


```
t-statistic: -0.916598531299853
p-value: 0.3721736112387992
```

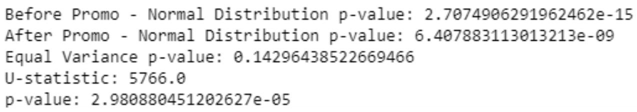
t-statistic value measures the difference between the means of two groups in terms of standard errors. In this case, we obtained a negative t-statistic value, typically indicating that the mean of the first group (sales before promotions) is lower than the mean of the second group (sales during promotions). The p-value being greater than the generally accepted significance level of 0.05 shows that the difference between the two groups is not statistically significant. Therefore, we can say that the promotion does not have a statistically significant effect on the sales quantities of products in the "Fast" category. The reason for this might be attributed to the small number of products, leading to an inability to capture the difference. The cause and consequence of this situation are illustrated by the graph mentioned above and the graph provided in the answer to question a. If we compare the "Medium" and "Slow" products, we obtain the following graphs and results respectively.







```
Before Promo - Normal Distribution p-value: 3.242049569962546e-05
After Promo - Normal Distribution p-value: 3.439827560214326e-05
Equal Variance p-value: 1.511528705149626e-05
U-statistic: 4030.0
p-value: 1.034100321581075e-30
```

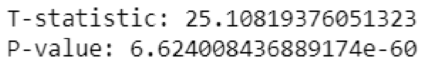


```
Before Promo - Normal Distribution p-value: 2.7074906291962462e-15
After Promo - Normal Distribution p-value: 6.407883113013213e-09
Equal Variance p-value: 0.14296438522669466
U-statistic: 5766.0
p-value: 2.980880451202627e-05
```

In both cases, as the assumptions were not met, the Mann-Whitney U test was used, and there is a difference between the Medium and Slow products before and during promotions. During the promotion, every product has higher weekly average sales compared to when there is no promotion.

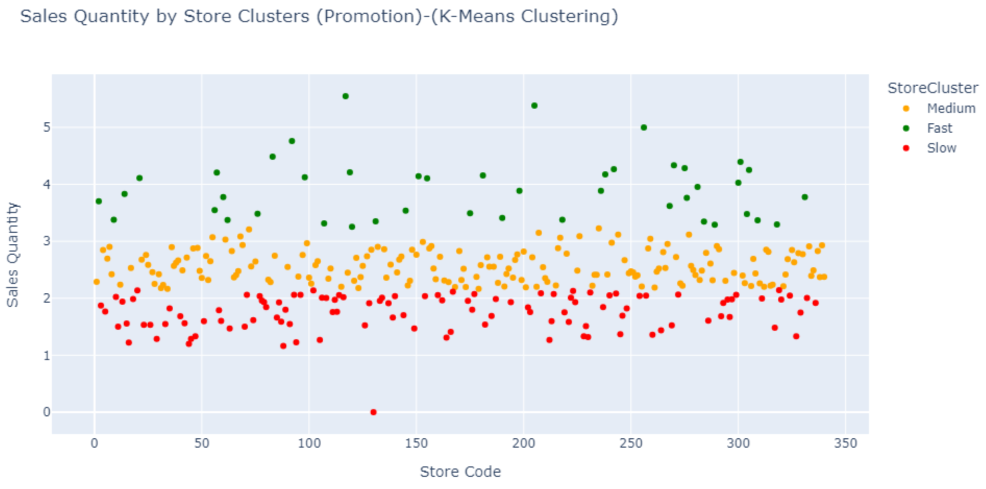
g. Is there any significant difference between promotion impacts of the Fast versus Slow stores?

The same situation applies to this question. Below, you can see the result of the requested situation.

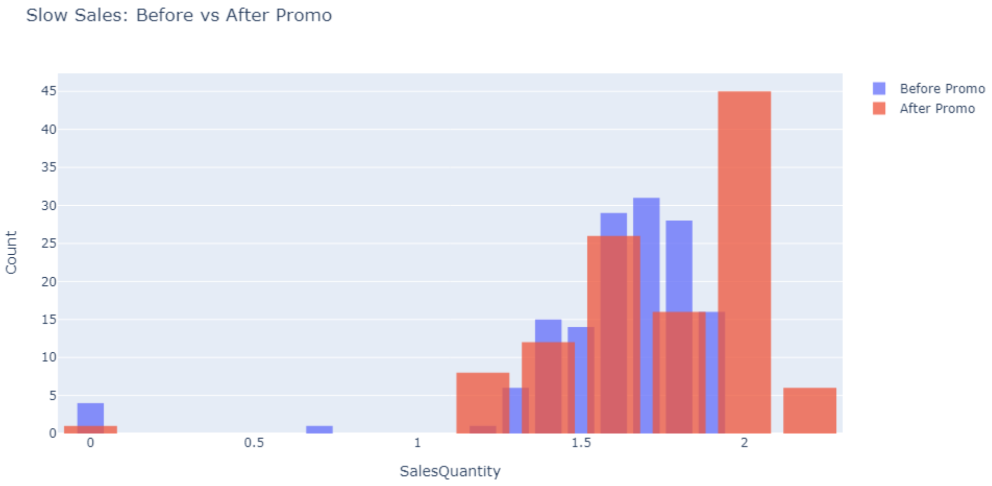


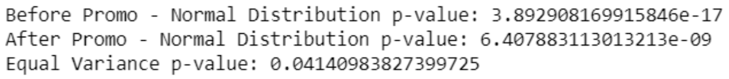
```
T-statistic: 25.10819376051323
P-value: 6.624008436889174e-60
```

The t-statistic value is 25.10819376051323, and the p-value is extremely small (6.624008436889174e-60). Considering the standard significance level (0.05), the p-value is significantly below this threshold. This tells us that there is a statistically significant difference between "Fast" and "Slow" stores during the promotional effects. The positive T-statistic indicates that "Fast" stores tend to have a higher average promotional effect compared to "Slow" stores. However, if we want to look at whether there is a statistically significant difference in the effects of Slow stores between the situation without promotion and the situation with promotion, which would be a better hypothesis, we must obtain the clustering graph in the promotional situation below.



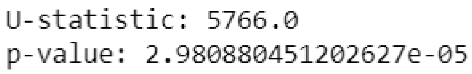
Before hypothesizing we must check the necessary assumptions.



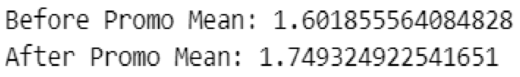


```
Before Promo - Normal Distribution p-value: 3.892908169915846e-17
After Promo - Normal Distribution p-value: 6.407883113013213e-09
Equal Variance p-value: 0.04140983827399725
```

Both the graphical results and test statistics reveal that the assumption of normal distribution is not met, indicating a left-skewed structure, and the Levene test result shows that the variances are not equal. Consequently, both the normal distribution and equal variance assumptions have been rejected. In this situation, a non-parametric test such as the Mann-Whitney U test should be applied instead of the t-test.

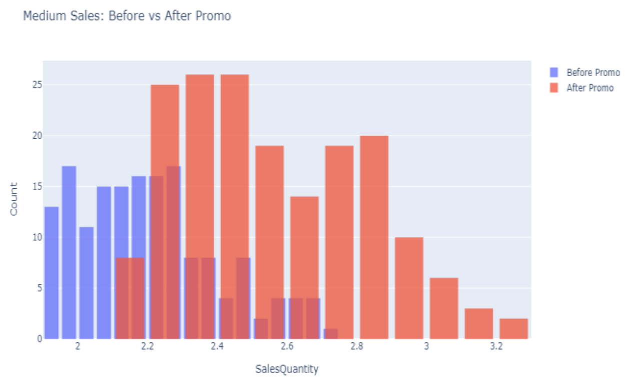


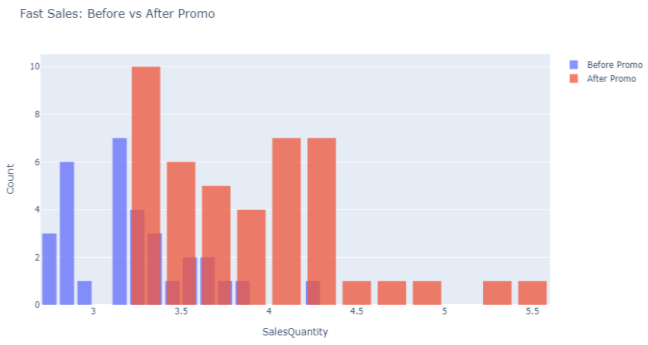
```
U-statistic: 5766.0
p-value: 2.980880451202627e-05
```

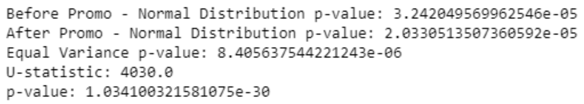


```
Before Promo Mean: 1.601855564084828
After Promo Mean: 1.749324922541651
```

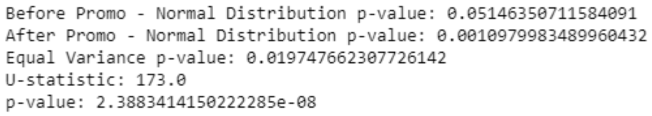
According to the results of the Mann-Whitney U test, we can say that there is a statistically significant difference between the pre-promotion and post-promotion sales of Slow products. The U-statistic provides a measure of the distribution of the samples, and in this case, it is 5766.0. The p-value is the probability of obtaining the observed statistic or a more extreme statistic, assuming the hypothesis is valid. In this case, the p-value is 2.980880451202627e-05, which is much lower than 0.05. This strongly indicates that there is a significant difference between the pre-promotion and post-promotion sales of Slow products. This suggests that the difference between the two groups (Without promotion and with promotion) is not by chance, meaning that the promotion has affected the sales of Slow products. This effect shows that the sales of Slow stores during promotion are higher than the sales of Slow stores without promotion. Likewise, we can access the results of Medium and Fast stores from the table below, respectively. However, it is crucial to conduct assumption checks.







```
Before Promo - Normal Distribution p-value: 3.242049569962546e-05
After Promo - Normal Distribution p-value: 2.0330513507360592e-05
Equal Variance p-value: 8.405637544221243e-06
U-statistic: 4030.0
p-value: 1.034100321581075e-30
```



```
Before Promo - Normal Distribution p-value: 0.05146350711584091
After Promo - Normal Distribution p-value: 0.0010979983489960432
Equal Variance p-value: 0.019747662307726142
U-statistic: 173.0
p-value: 2.3883414150222285e-08
```

In general, we can say that sales effects increase when all kinds of stores have promotions.

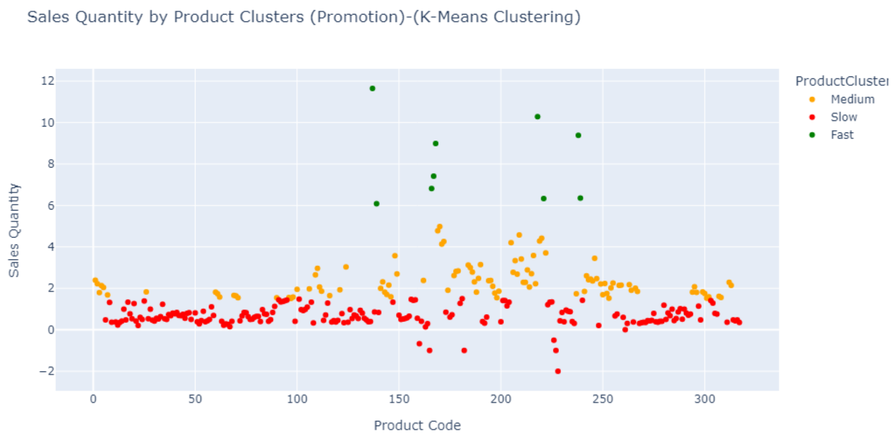
3. Data = Assignment4.1b.csv

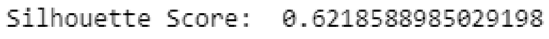
The data contains daily sale of sample items in several stores after the initial data ends. This will enable you to evaluate your models predictive power on the newly observed data.

B. You are asked to measure how well your model have worked on this new data. Based on the model developed in part A forecast what would the effect of promotion 5 will be on sample store – item pairs. Compare your forecasts results for promotion5 with the real observed sales during that period.

What measure would you use for goodness of fit?

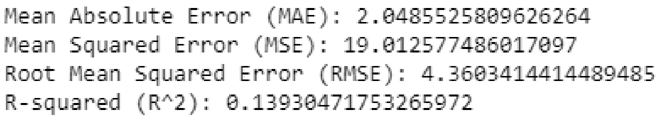
"Goodness of fit" is a measure of how well a statistical model predicts a data set. In Regression Analysis, R-Squared or Adjusted R-Squared values are used. Metrics like MSE (Mean Squared Error), MAE (Mean Absolute Error), and RMSE (Root Mean Square Error) are used in model formation. These metric values show the average difference between the actual values and the model's predictions. In Logistic Regression or Other Classification Models, classification metrics such as Precision, Recall, F1 Score, and Area Under the Receiver Operating Characteristic (AUC-ROC) are used as Confusion Matrix metric values. They show the number of correct and incorrect predictions of the model. In statistical tests, the Chi-Square goodness of fit test is used. In Clustering Analysis, a metric value showing how well the clusters are separated is used, called the Silhouette Score. It calculates a cohesion and separation value for each sample and looks at the difference between these values. It takes values between -1 and 1, and higher values indicate better clustering. Since our first established model is a clustering analysis, we need to use the Silhouette Score. Also, since we have set up a regression model and used linear regression and random forest algorithms for our new model, we will compare the RMSE, MAE, MSE, and R-squared values.





```
Silhouette Score:  0.6218588985029198
```

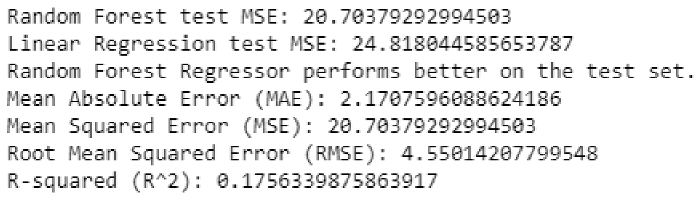
The Silhouette Score for Promotion 5 was calculated as 0.6218588985029198. Since this value is between -1 and 1, the fact that this score is positive and higher than 0.5 generally indicates that the clustering has been well-formed. However, different parameters or clustering techniques can be applied to improve clustering or find a more accurate model. The same holds true for predicting sales quantities.



```
Mean Absolute Error (MAE): 2.0485525809626264
Mean Squared Error (MSE): 19.012577486017097
Root Mean Squared Error (RMSE): 4.3603414414489485
R-squared (R^2): 0.13930471753265972
```

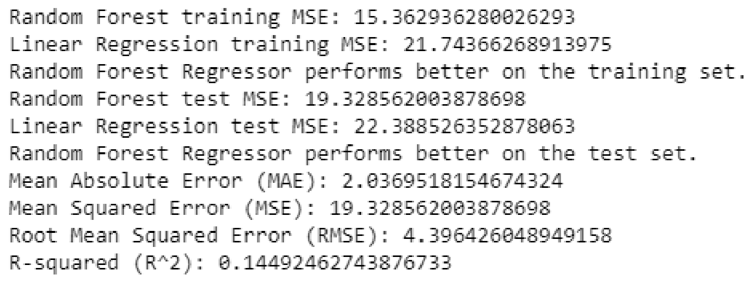
How good is your model developed in step 1?

Step 1



```
Random Forest test MSE: 20.70379292994503
Linear Regression test MSE: 24.818044585653787
Random Forest Regressor performs better on the test set.
Mean Absolute Error (MAE): 2.1707596088624186
Mean Squared Error (MSE): 20.70379292994503
Root Mean Squared Error (RMSE): 4.55014207799548
R-squared (R^2): 0.1756339875863917
```

Step 2



```
Random Forest training MSE: 15.362936280026293
Linear Regression training MSE: 21.74366268913975
Random Forest Regressor performs better on the training set.
Random Forest test MSE: 19.328562003878698
Linear Regression test MSE: 22.388526352878063
Random Forest Regressor performs better on the test set.
Mean Absolute Error (MAE): 2.0369518154674324
Mean Squared Error (MSE): 19.328562003878698
Root Mean Squared Error (RMSE): 4.396426048949158
R-squared (R^2): 0.14492462743876733
```

In Step 2, the model performs better according to the MAE, MSE, and RMSE metrics, while the R-squared value is slightly lower compared to the model in Step 1. In both cases, the values are relatively low, and it could be beneficial to further develop the model to better explain the variance in the data.

What are the main problem points causing bad fits?

The situations that can lead to poor performance of the model include data quality, the relationship of independent variables with the target variable, the values applied to hyperparameters, overfitting or underfitting situations, and most importantly, the appropriateness of the model. Not fully understanding the desired condition affects the accuracy and success of the model I have built.

What would you change in step 1?

To improve the performance of Step 1, first, a review of the features used may be needed. By using only the 'StoreCode', 'ProductCode', and 'Promotion' features, we may be insufficient in fully explaining all the variance in sales quantity. Adding other features in the dataset or different explanatory variables and checking their impact on the model could be beneficial. Second, we can review the data preprocessing process. Steps like feature scaling, categorical variable encoding, and handling of outlier values (through transformation techniques or robust methods) may enhance the model's performance. Third, we could expand the choice of the model used. Different regression models (such as Gradient Boosting, XGBoost, SVR, or neural network) could be tried to find the one that provides the best result. Fourth, we can adjust the hyperparameters of the model. For the Random Forest model, optimizing hyperparameters like 'n_estimators', 'max_depth', 'min_samples_split' could improve the model's performance. Lastly, we could use cross-validation in the model training process. This allows us to predict the model's performance on unseen data more accurately. These changes could be effective in enhancing the model's performance, but the effectiveness of each change will depend on the dataset.

Note: You will not be judged on goodness of your fit at step 1.

4. Report:

Please write a small report covering the points below. You can assume that client`s goal is to understand the reaction of the products and stores during promotions and to measure the effect of promotions on sales in the store. Please include your code at the end of the report. We would like to know how you achieve the conclusion. Which methods did you use?

- Make recommendations.
- Show your work by providing code.
- Report statistics for all models used.
- Interpret results.
- **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?**

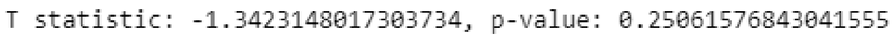
The weekly sales of each store can vary significantly based on factors such as store traffic, local demand, and other influences. These are general criteria and may not be applicable in every case. For instance, in some situations, sales values instead of sales quantities might be used to categorize "Fast", "Medium", and "Slow" selling products. Also, the criteria used to define these categories may differ across various sectors and business models, depending on business needs and the specific requirements of a situation.

Additionally, certain variables or data structures that may be essential could be added to the dataset. The seasonality factor could be a decisive variable, as there may be spikes in sales during specific times of the year (such as summer discounts, winter sales, holiday seasons, etc.). Therefore, including seasonal effects in sales data could be advantageous. One of the most critical conditions is the pricing of products. A product's price generally directly affects its sales. If pricing data is available, it could be a significant variable in the developed models. Advertising and marketing campaigns typically impact sales. If information about where or when advertising occurred is recorded, more successful results could be obtained. Competitor product prices and presence can also influence sales. If access to this data is possible, the effect of competitor products on sales can be analyzed. Moreover, different product categories may have distinct sales trends. For example, food products are often sold more consistently, whereas the sales of luxury goods may be more volatile. If product information can be added to the dataset, more detailed studies can be carried out. Finally, knowing the demographic information of customers (such as age, gender, income level) may affect sales.

Bonus

- **Is there any significant difference in item return rates after promotions?**

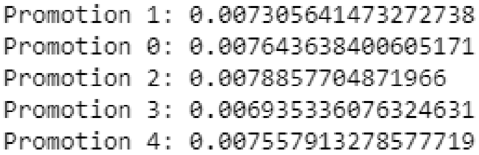
Three different scenarios were considered for this condition. In the first scenario, the situation of whether there is a statistically significant difference between the ratio of products returned to the products sold according to promotions was examined.



```
T statistic: -1.3423148017303734, p-value: 0.25061576843041555
```

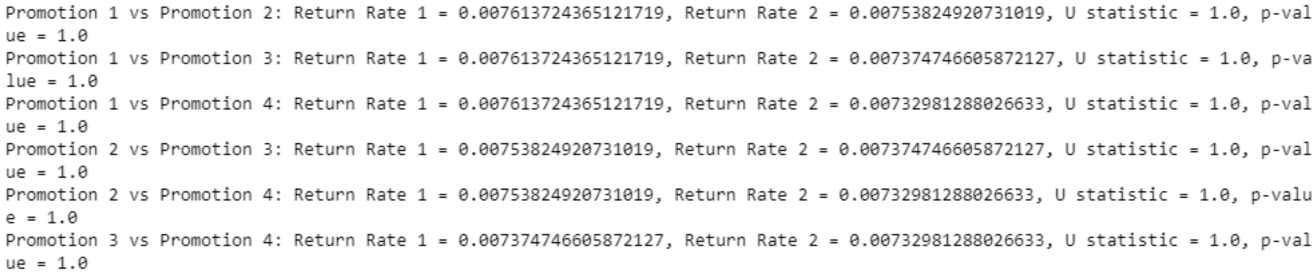
When the results are analyzed, the t-statistic is -1.34 and the p-value is 0.25. In general, if the p-value is greater than 0.05, we accept that the null hypothesis cannot be rejected. In this case, since the p-value is greater than 0.05, we can conclude that there is no statistically significant difference between the number of returns and the ratio of the number of sales. This means that promotions do not have a significant effect on return rates.

In the second case, we calculated the return rate for the non-promotion period and the return rates for each promotion code. Then, the ratio of the calculated values before and after the promotion was compared with the Mann-Whitney U test. Mann-Whitney U test statistic: 12.0, p-value: 0.34285714285714286 and there was no significant difference between the two groups (p-value greater than 0.05). This means that the return rates after the promotions did not change significantly compared to before the promotions.



```
Promotion 1: 0.007305641473272738
Promotion 0: 0.007643638400605171
Promotion 2: 0.0078857704871966
Promotion 3: 0.006935336076324631
Promotion 4: 0.007557913278577719
```

Finally, the subsequent return and sales amounts were calculated for the relevant promotion periods (1,2,3,4). With the help of these values, the values of the calculated return rates were determined. Then, Mann-Whitney U test was applied between all promotion pairs. As a result, a significant difference in product return rates after the relevant promotions was checked and the following results were obtained.



```
Promotion 1 vs Promotion 2: Return Rate 1 = 0.007613724365121719, Return Rate 2 = 0.00753824920731019, U statistic = 1.0, p-val
ue = 1.0
Promotion 1 vs Promotion 3: Return Rate 1 = 0.007613724365121719, Return Rate 2 = 0.007374746605872127, U statistic = 1.0, p-va
lue = 1.0
Promotion 1 vs Promotion 4: Return Rate 1 = 0.007613724365121719, Return Rate 2 = 0.00732981288026633, U statistic = 1.0, p-val
ue = 1.0
Promotion 2 vs Promotion 3: Return Rate 1 = 0.00753824920731019, Return Rate 2 = 0.007374746605872127, U statistic = 1.0, p-val
ue = 1.0
Promotion 2 vs Promotion 4: Return Rate 1 = 0.00753824920731019, Return Rate 2 = 0.00732981288026633, U statistic = 1.0, p-valu
e = 1.0
Promotion 3 vs Promotion 4: Return Rate 1 = 0.007374746605872127, Return Rate 2 = 0.00732981288026633, U statistic = 1.0, p-val
ue = 1.0
```

The results show that there is no significant difference in return rates across all promotions (p-values 1). This means that different promotions have similar effects, at least in terms of return rates. The return rates of each promotion are close and similar to each other.