Gayathri Gopalan

Retail Sales Marketing Analysis

Overview

This detailed report is intended to examine a retail dataset and to determine whether or not the sales of outdoor products are growing, maturing or declining. It also projects the analysis to prospective future sales with past performance based on the results obtained.

Metadata

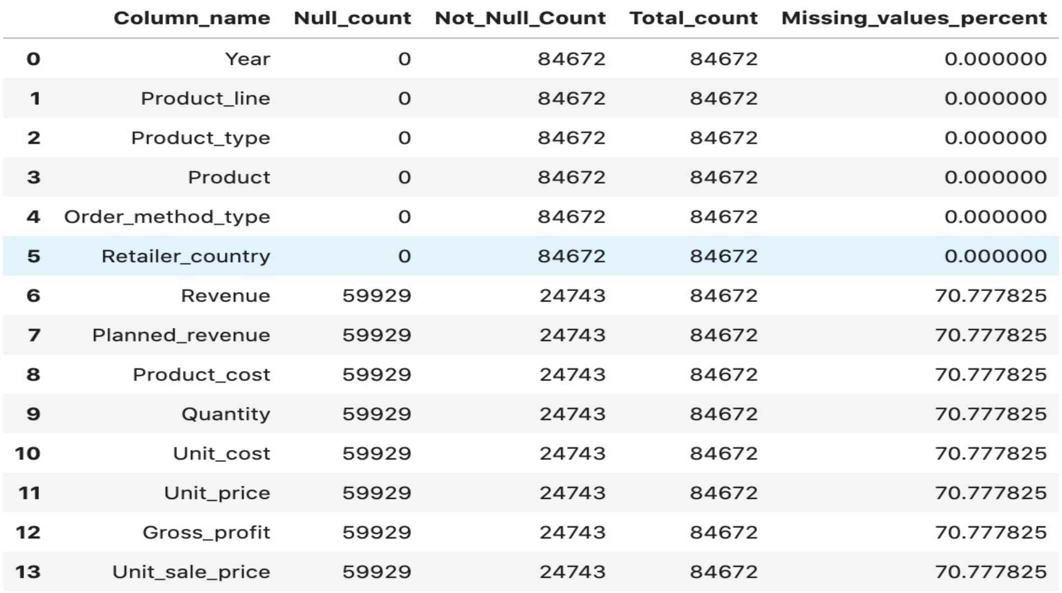
The given retail dataset comprised of the following columns, which will eventually be analyzed and based on their statistical significance to the target variable, they may be converted into features accordingly. The columns are as follows

- *Year* – Year of Sales
- *Product line and type* – Category of product sold (eg: Camping Equipment)
- *Order method type* – Mode of order (eg: Telephone, email)
- *Retailer country* – Country of Sales
- *Revenue, planed revenue* – Actual Revenue and Anticipated Revenue of Product
- *Product Cost* – Cost price of the product sold
- *Quantity* – Quantity of Sales
- *Unit cost* – Cost per unit/product
- *Unit price* - Sales per unit/product
- ***Gross profit*** – Yield
- *Unit sale price* – Sales on discount per unit/product

Considering retail sales dataset, we can clearly identify that the success is determined based on the 'Gross Profit' or yield. We will use the sale for performing various machine learning models to determine and predict the future sales.

Data Cleaning

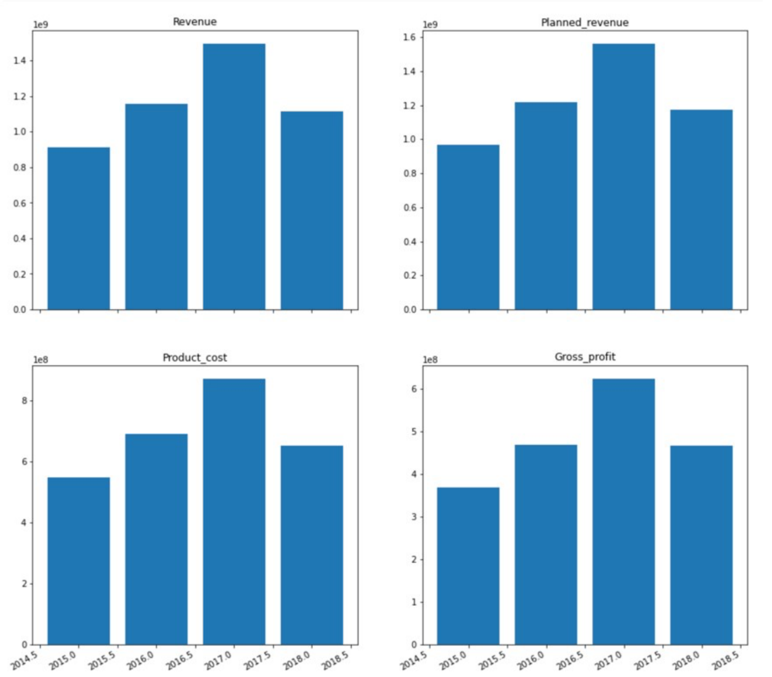
Retail Sales dataset originally is comprised of 84672 rows and 14 columns, but we are not sure about the missing partitions count and how the values in each of them are comprised. Performing a missing value analysis, we can see the following results

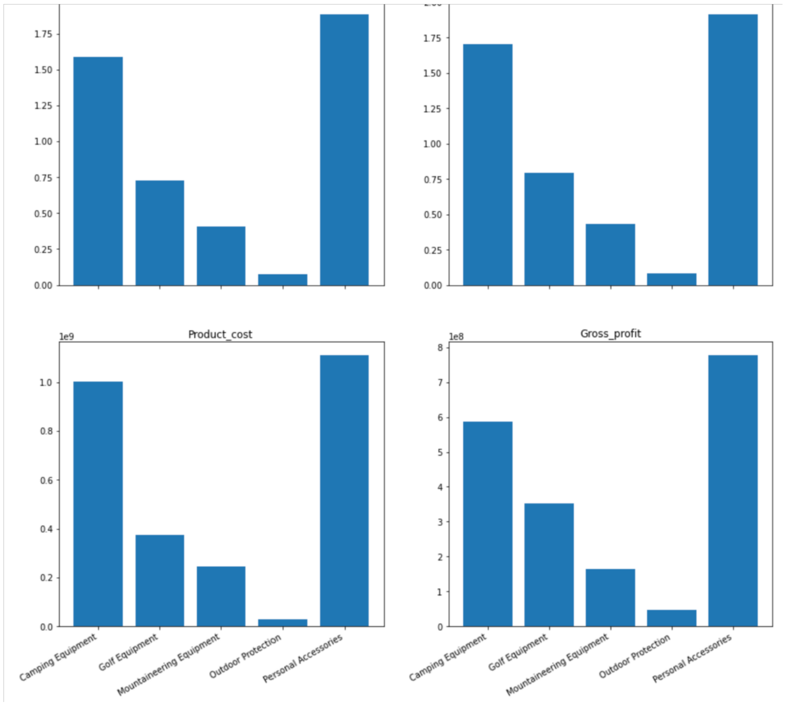


	Column_name	Null_count	Not_Null_Count	Total_count	Missing_values_percent
0	Year	0	84672	84672	0.000000
1	Product_line	0	84672	84672	0.000000
2	Product_type	0	84672	84672	0.000000
3	Product	0	84672	84672	0.000000
4	Order_method_type	0	84672	84672	0.000000
5	Retailer_country	0	84672	84672	0.000000
6	Revenue	59929	24743	84672	70.777825
7	Planned_revenue	59929	24743	84672	70.777825
8	Product_cost	59929	24743	84672	70.777825
9	Quantity	59929	24743	84672	70.777825
10	Unit_cost	59929	24743	84672	70.777825
11	Unit_price	59929	24743	84672	70.777825
12	Gross_profit	59929	24743	84672	70.777825
13	Unit_sale_price	59929	24743	84672	70.777825

Missing Value Analysis of Retail Dataset

This summary shows that there are 70% missing values in some of the key metrics like Revenue, Planned Revenue, Cost, Quantity, Profit. We have to eventually delete the data, but let us see which subgroup has maximum no. of missing values. This would educate us on how we can deal with these missing values (whether to delete them or fill them with zero or mean and so on)





Distribution of Null Values Across Different Categories shows that they have a distributed percentage of missing values. Therefore, we can confirm that we have no choice but to get rid of them. 70% of data removal is bad, but if we are manipulating 70% of the data with the mean of rest of the 30% - it would lead to worse results. Therefore, we decide to remove the 70% of missing values.

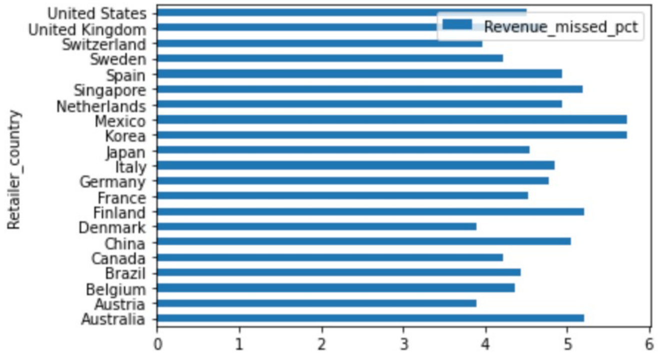
We also see some of the whitespaces and '- '(hyphens) on the key metrics like Revenue. We have to remove them in order to feed to the model. And also, there are negative values represented in "(value)" instead of "-value" – we remove them too.

Exploratory Analysis

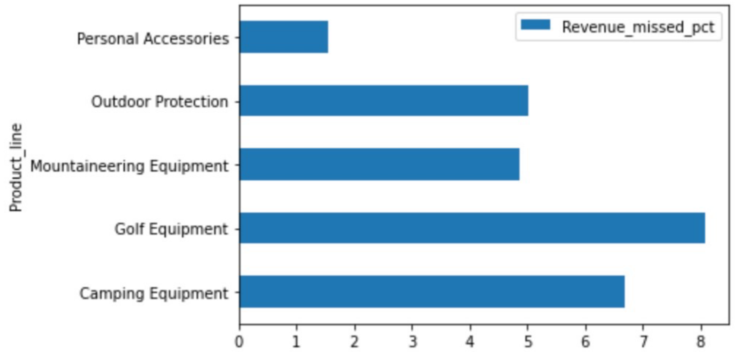
Now that we have understood the dataset in our consideration and cleaned it, let us explore the various dimensions and features of this dataset and try to understand if we can derive some conclusions from it.

Let us create a column called ‘Revenue missed percentage’ which will be used to find the difference percentage from actual revenue and planned revenue

```
  results['Revenue_missed_pct'] =
((results['Planned_revenue']results['Revenue'])/(results['Plann
ed_revenue']))*100.0
```

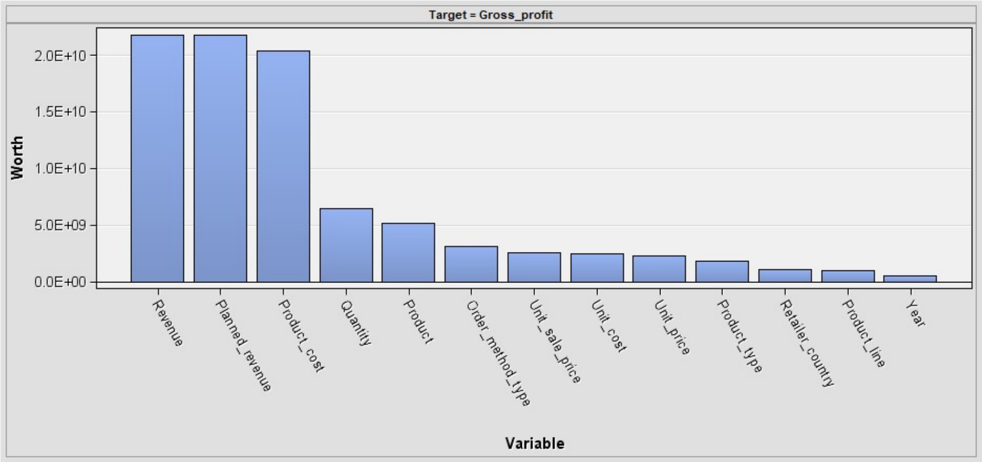


These figures shows that the revenue miss is distributed across different dimensions and not concentrated towards one country/product line

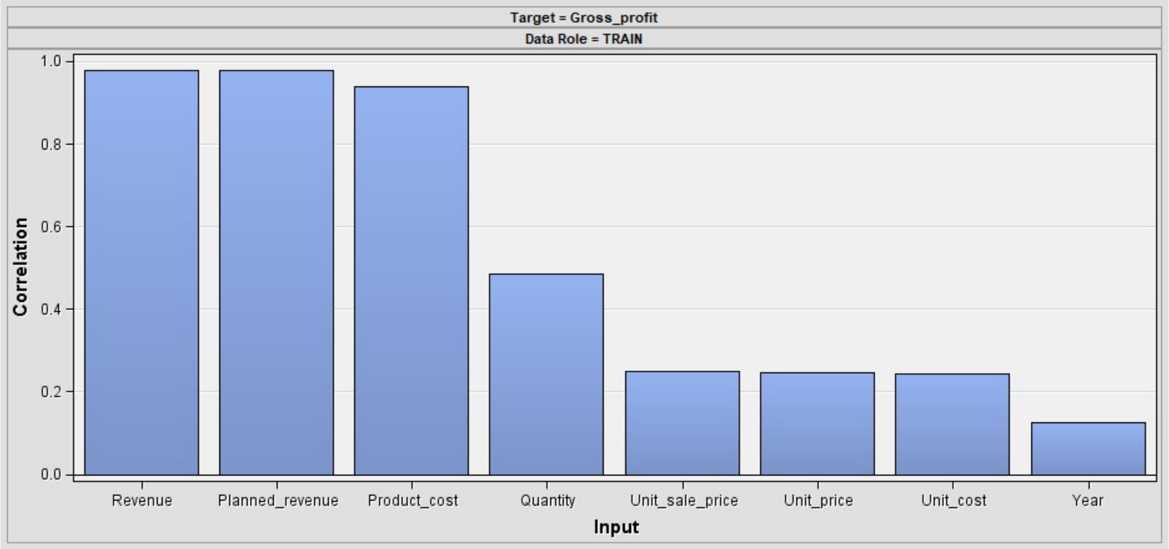


Descriptive Analysis

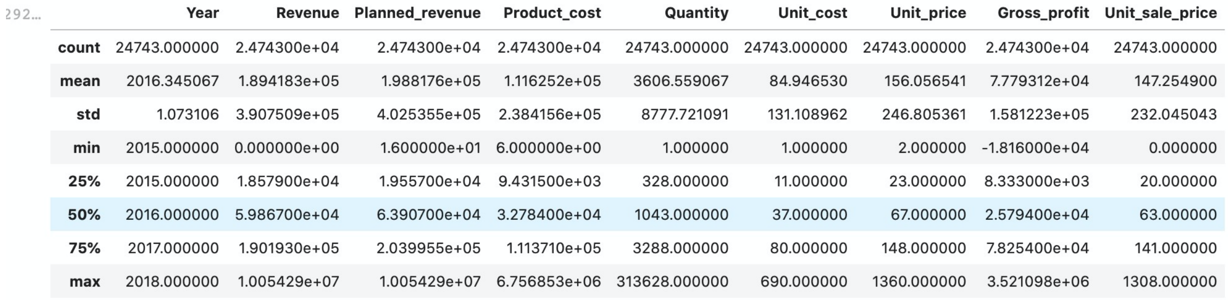
Now that we have explored the data let us look at the minor details about the mean, standard deviation distribution and so on



Accessing the worth of the variable against the target variable which is the Gross Profit we can see that three features standout. And a similar pattern is seen in the correlation between the datasets



Let us take a look at the statistical numbers of these metrics and determine the distribution of these dataset



292…	Year	Revenue	Planned_revenue	Product_cost	Quantity	Unit_cost	Unit_price	Gross_profit	Unit_sale_price
count	24743.000000	2.474300e+04	2.474300e+04	2.474300e+04	24743.000000	24743.000000	24743.000000	2.474300e+04	24743.000000
mean	2016.345067	1.894183e+05	1.988176e+05	1.116252e+05	3606.559067	84.946530	156.056541	7.779312e+04	147.254900
std	1.073106	3.907509e+05	4.025355e+05	2.384156e+05	8777.721091	131.108962	246.805361	1.581223e+05	232.045043
min	2015.000000	0.000000e+00	1.600000e+01	6.000000e+00	1.000000	1.000000	2.000000	-1.816000e+04	0.000000
25%	2015.000000	1.857900e+04	1.955700e+04	9.431500e+03	328.000000	11.000000	23.000000	8.333000e+03	20.000000
50%	2016.000000	5.986700e+04	6.390700e+04	3.278400e+04	1043.000000	37.000000	67.000000	2.579400e+04	63.000000
75%	2017.000000	1.901930e+05	2.039955e+05	1.113710e+05	3288.000000	80.000000	148.000000	7.825400e+04	141.000000
max	2018.000000	1.005429e+07	1.005429e+07	6.756853e+06	313628.000000	690.000000	1360.000000	3.521098e+06	1308.000000

We can see the data is distributed across the different years (2015, 2016, 2017, 2018). Each year data is fitting into each quantile which means that there is an even distribution of data points across the years. And we do not see much surprise when looking at standard deviations and means

Feature Engineering

We can see that from our previous correlations and worth graph, we are going with three main features

- Planned Revenue
- Revenue
- Product Cost

And our target variable is ‘Gross Profit’. To deliver a strategic plan, let us think about a new feature that will be used in our logistic regression

```
df['Missed_Revenue']=df['Planned_revenue']-df['Revenue']
df['revenue_to_cost_rate']=(df['Revenue']/df['Product_cost'])
```

Revenue to cost Rate would be used to determine the worthiness of the product like ratio of cost to sales. And we will also convert them to a Boolean value called ‘Product_satisfied’ – this will be determined by splitting on the revenue_to_cost_rate’s mean

```
df['profit_satisfied'] = np.where(df['revenue_to_cost_rate']>1.9, 1, 0)
```

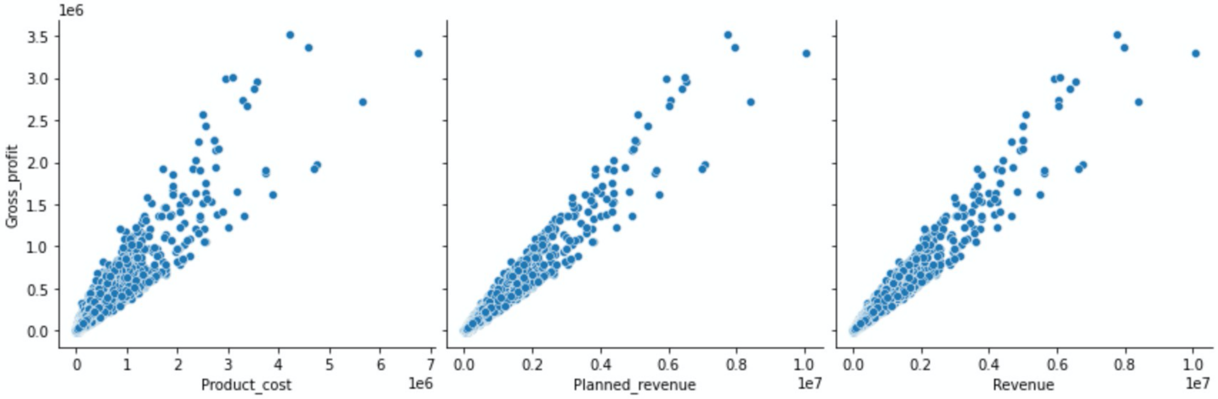
Predictive Analysis

1. Multi Variable Linear Regression

For our model the following are the key parameters to be considered.

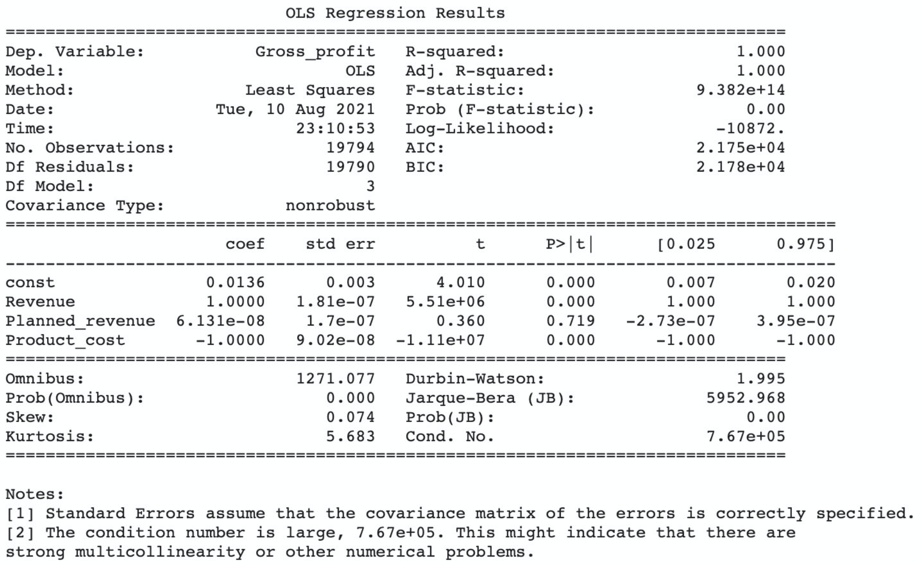
Independent Value – Gross_profit
Dependent Value – Planned Revenue, Revenue and Product Cost

Just to understand that there are no outliers with respect to the independent variable, we see the following pattern



We can see that the pattern is pretty linear and there are no outliers, the dependency is pretty linear (which is encouraging).

When we run a linear regression model, we see the following and let us then interpret the method


```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Gross_profit   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 9.382e+14
Date:                Tue, 10 Aug 2021   Prob (F-statistic):               0.00
Time:                        23:10:53   Log-Likelihood:                -10872.
No. Observations:               19794   AIC:                         2.175e+04
Df Residuals:                   19790   BIC:                         2.178e+04
Df Model:                           3
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
const               0.0136      0.003      4.010      0.000       0.007       0.020
Revenue             1.0000   1.81e-07   5.51e+06      0.000       1.000       1.000
Planned_revenue  6.131e-08    1.7e-07      0.360      0.719   -2.73e-07    3.95e-07
Product_cost       -1.0000   9.02e-08  -1.11e+07      0.000      -1.000      -1.000
==============================================================================
Omnibus:                     1271.077   Durbin-Watson:                   1.995
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5952.968
Skew:                           0.074   Prob(JB):                         0.00
Kurtosis:                       5.683   Cond. No.                     7.67e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.67e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Here are the key metrics to interpret from the above results

Coefficient Interpretation:

- **Revenue:** The coefficient for Revenue is 1.0000, with a very low p value (0.00<0.05) The coefficient is positive, so it is positively correlated with Profit (Independent/ target variable) p value is statistically significant.
- **Planned_revenue:** The coefficient for Planned_revenue is 6.131e-08, with a high p value (0.719 >0.05) The coefficient is positive, so it is positively correlated with Profit (Independent/ target variable) p value is not statistically significant.

- **Product_cost:** The coefficient for Product_cost is -1.0000, with a very low p value (0.00<0.05) It makes sense, as if there is increase in cost, the profits will decrease. SO it is negatively correlated. p value is statistically significant.

R - squared:

- R-Square is 1 meaning that 100% of the variance in Profit is explained by the 3 factors mentioned above
- This is a perfect R-squared value, which implies that independent variables andd ddependent variables are having a strong correlation.

F statistic:

- F-value has a very low p value (practically low). Meaning that the model fit is statistically significant, and the explained variance isn't purely by chance.

The fit is significant. Let us visualize how well the model fit the data. From our parameters, the linear regression equation is:

*Gross_Profit= 0.0136 + 1.0000 Revenue + 6.131e-08 Planned_revenue -1.0000 *Product_cost*

Cross Validation

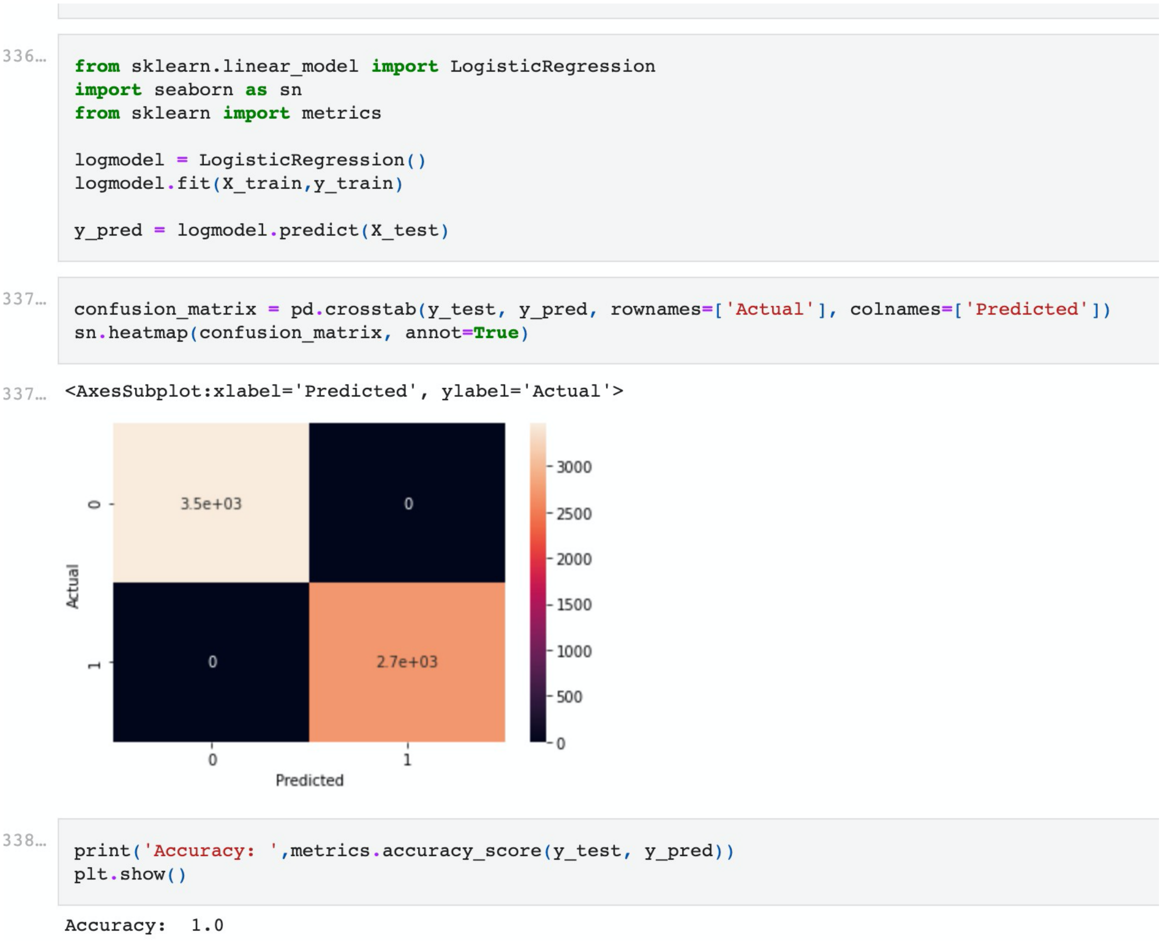
After validating the results with the rest of the 20% of the test data, we see very good results with minimal error rate

The model has an R-squared error of 0.999 which is a great performance to the model

2. Logistic Regression

Considering the logistic regression on profit satisfied, which is the derived column from revenue_to_cost_ratio, we have the following hyper parameters

- Target - profit_satisfied
- Dependent Values - Revenue, Planned Revenue, Product Cost



336…

```
from sklearn.linear_model import LogisticRegression
import seaborn as sn
from sklearn import metrics

logmodel = LogisticRegression()
logmodel.fit(X_train,y_train)

y_pred = logmodel.predict(X_test)
```

337…

```
confusion_matrix = pd.crosstab(y_test, y_pred, rownames=['Actual'], colnames=['Predicted'])
sn.heatmap(confusion_matrix, annot=True)
```

337…

```
<AxesSubplot:xlabel='Predicted', ylabel='Actual'>
```

338…

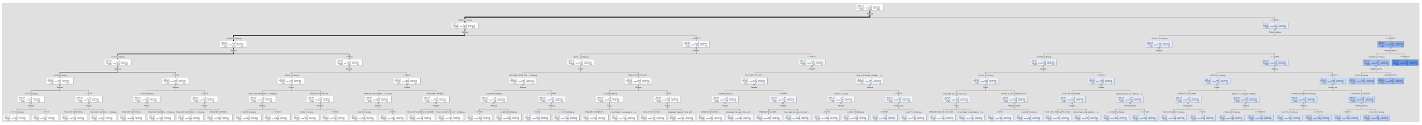
```
print('Accuracy: ',metrics.accuracy_score(y_test, y_pred))
plt.show()
```

```
Accuracy:  1.0
```

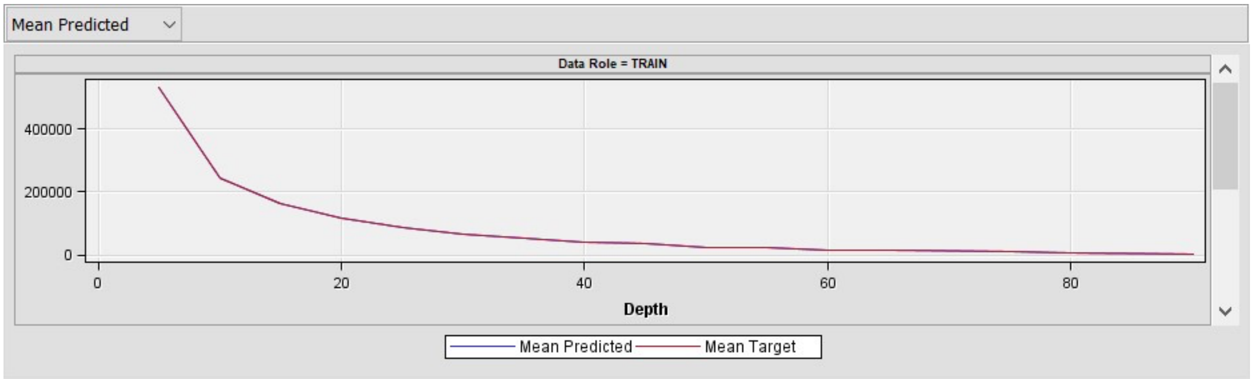
Model could predict the values very accurately with no false negative and false positives. This means that our parameters are on target. This is confirmed by cross validating with 20% of the data.

3. Decision Tree

When we run the decision tree with a decision on the profit, across the three features (revenue, Planned Revenue, Product Cost) – we see the following tree, which has been configured for 5degree depth



We can see that the predicted and target values are in the same direction at different level of depth



Recommendations

After detailed analysis we can say that the sales of outdoor products world-wide is **growing** with the increase in the investments.

When we invest more on the products, we will have good yield. So we recommend all products across worldwide to increase the investment and they can yield a growth ratio of 1.9 in average (based on the existing data).

Learnings & Future Implementation

Some of the shortcomings of this analysis are

- Too many missing values
- More features on the data like customer related info, months, day, seasonality information
- Breakdown of cost and sales into multiple factors (operational cost, labour cost and so on)
- With more transaction-level customer metrics, we can build recommendation systems

If we can gather these details, we can expand our analysis to different dimensions.