SAMPLE SUPERSTORE

Statistical Analysis on Sales Data

Name: Deenadhayalan Ravi
Department: Computer Science
Organisation: City University of London
email:deenadhayalan.ravi@city.ac.uk

Abstract— **The study has been conducted on the sales data of Sample Superstore with order details, customer details, product details and business values. Some business insights and analyzing the factors which influences the profit of the order will be the main focus of analysis using descriptive statistical analysis and inferential statistical analysis – Multiple Linear Regression models.**

Keywords—Multiple Linear Regression, k-Fold cross validation, Sample Superstore EDA

I. INTRODUCTION

In this modern era with all the technology advancement Data Analysis becomes an important tool for running a successful business. It can be used to refine, streamline and improve each and every area of the business.

Data Analysis can formulate the sales strategies of the companies to achieve the mission and vision of the company. The analysis can be from product perspective where the significance of their products in the market can be studied and market share, wallet share of their products can be analysed. From the customer front, analysis can be done on the customer segmentation, customers buying behaviour, customer satisfactions or dis-satisfaction with the customer satisfactory attributes, lost customers and a strategy how to regain them. From the competitor's side, analysis can be done on the competitors dynamics which includes the competitors market share, product/technology comparison. Data Analysis gives you insights on pricing strategy, budgeting, forecasting, how to increase the business turnover and profits, how to minimize the losses.

In this paper we will be performing both descriptive statistical analysis and inferential statistical analysis on the Sample Superstore dataset in order to identify significant insights from four years of sales data and support the decision making of the business.

II. RESEARCH QUESTIONS

The top management of the company is interested in analysing the past four years data and formulate strategies to excel in the market, here are some important research questions for analysis,

A. How is the total sales and profits spreaded over the following attributes,

1. Over the 4 years(2014-217)
2. Customer Segmentation
3. Geographical segregation
4. Category or sub-category

The explanation to the above questions should give the overview of the business and highlight the important points which will be helpful for decision making.

B. Which customers contributes to the high business value and what are the products sold to them?

C. Which customers contributes to the loss and what are the products sold to them?

D. Create a model to predict the profit of the order. Which are the attributes that contributes in deciding profits?

III. DATA (MATERIALS)

A. *Key Characteristics*

Our data is all about 4 years sales of the products in the superstore which has 21 attributes with 9994 instances, each row is a product line item of the order. This dataset is a mixture of quantitative and qualitative variables.

Quantitative Variable: 'Sales' is a continuous attribute which is the total sale value of the product (i.e., unit sale price * order quantity). 'Profit' is a key continuous attribute which shows the profit/loss of the product in the order. 'Discount' is the discrete attribute which shows the discount to the product in the order.

Qualitative Variables: The nominal attributes in this dataset gives information like shipment mode, customer segment, customer location, product category and product name and the date of order and shipment is the only ordinal attribute in the dataset.

B. *Is It Suitable for Our Study?*

Basic data pre-processing has to be done for the complete dataset which will be explained in the upcoming slides and attributes like 'Sales', 'Profit', 'Order Date', 'Region', 'States', 'Category', 'Sub-Category', 'Segment' are good enough to answer the question II-A, II-B and II-C. Some more attributes have to be manipulated and derived in order to explain the question II-D.

C. *Data Limitations and Assumptions*

1. There is a limitation on the dataset that there are no much features which explains the product.

2. Unit Price of each and every product is same across the four-year sales which is practically not possible

IV. ANALYSIS

Our research questions can be segregated into descriptive statistical analysis and inferential statistical analysis.

A. *Data Preparation:*

The following pre-processing are done in the dataset for our analysis purpose,

1. Check for null values and duplicate instances and removed, if any.
2. Droping the columns which are either non-value added or with duplicated information i.e., ‘Customer ID’, ‘Postal Code’, ‘Country’, ‘Product ID’.
3. Aligning the date and time format of ‘Order Date’ and ‘Shipment Date’ columns.
4. By manipulating the given continuous variables, arrived the following variables,
   Unit_sales_price = Sales / Quantity
   Unit_quotation_price = (Sales/Quantity) / (1-Discount)
   Unit_production_cost = (Sales – Profit) / Quantity
   Profit%= ((Unit_sales_price-Unit_production_cost)/ Unit_sales_price) * 100
5. Extracting years, months weekdays from the date stamp of Order ID to separate 3 columns as order_year, month_of_order_year and weekday_of_order_year respectively.

The following pre-processing are done for the modeling purpose - multiple linear regression (ref step 7 in the Jupiter Notebook),

6. Extracting the quarters of the year, which may have an impact on the target attribute.
7. Categorizing the order ID based on no. of line items in each order, which may have an impact on the target attribute
8. Using label encoding to convert all the useful categorical attributes to numerical attributes for correlation analysis.

B. *Data Derivation:*

All the descriptive analysis are based on the central limit theorem and in this section, we will be analysing all our research question except the last one.

1. Order year Vs Total Sale Value and Profit:

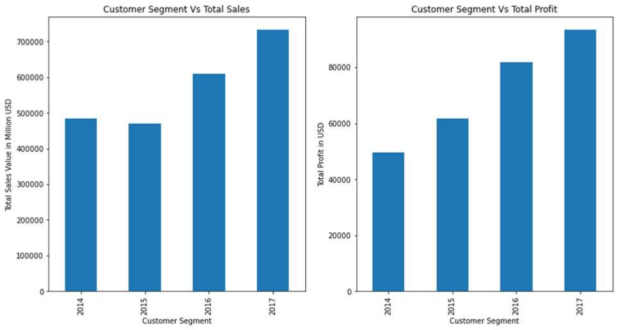


Fig:1

Inference: The total sale value and the total profit has gradually raised over the years from 2014 and reached more than 700,000 USD and more than 90,000 USD respectively in 2017.

2. Analysing the monthly total sales for each year:

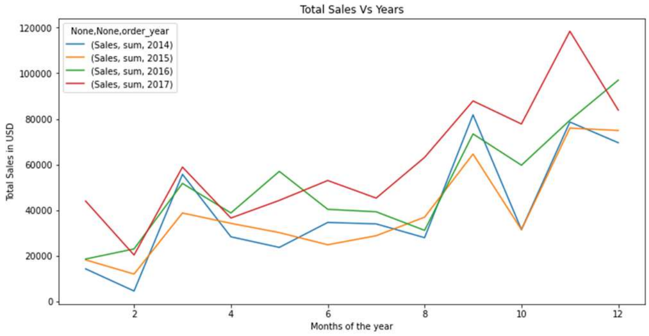


Fig:2

Inference: The peak sale of the year is at November and December for all the years. And there is notable raise in September followed by March.

3. Analysing the monthly total profit for each year:

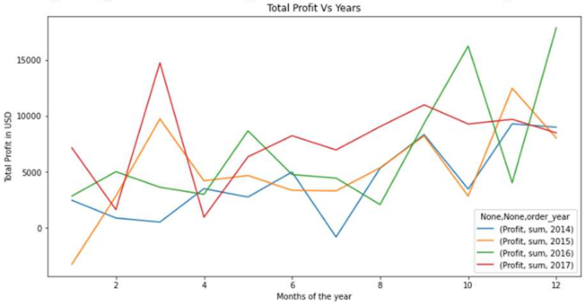


Fig: 3

Inference: The peak profiting months are, for 2014 - Nov & Dec, 2015 - Nov followed by Mar, 2016 – Dec followed by Oct and 2017 - Mar followed by Sep.

4. Analysing Customer Segment with Sales and Profits:

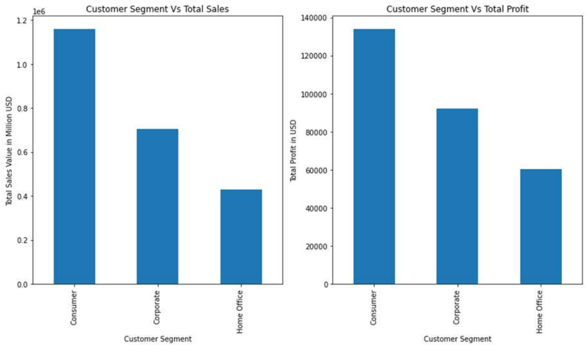


Fig: 4

Inference: Consumer segment is toping both the tables with total sales over 1 Million USD and profit more than 130,000 USD.

5. Analysing Customer by region with Sales and Profits:

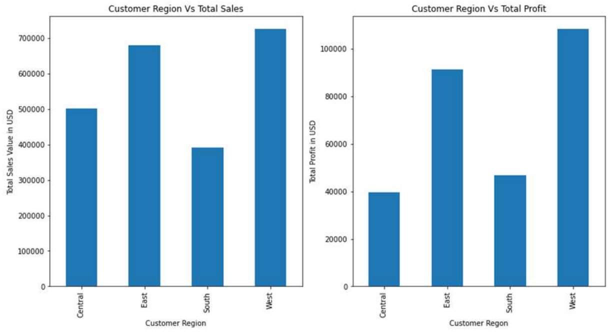


Fig: 5

Inference: Both sales and profits are good in West followed by Eastern Region. The notable point here is even though the total sales in Central is higher than South, their profits are lesser than South, which means there is a loss-making business happened in Central region. Further analysis on Central will give a better clarity.

6. Diving one step deeper let's analyse the states – Sales and Profit/loss:

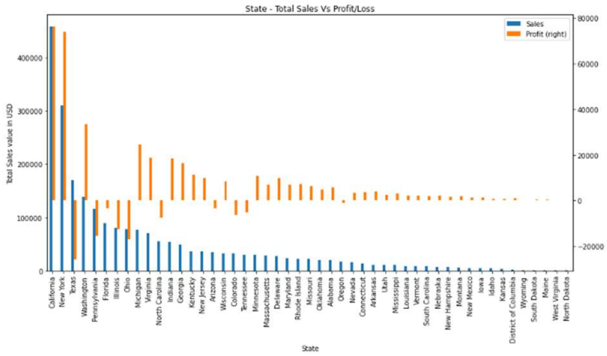


Fig: 6

Note: The profit is layered in the secondary Y axis.

Inference: by analysing the top selling states and their profits, California (West), New York (East), Washington (West) states are having higher overall sales and their profits are also on the positive side, whereas Texas (Central), Illinois (Central) states are having higher sales but with negative margins (loss business). Both of these points answer the phenomena in regional analysis.

7. Analysing the Product Category Vs Sales and Profit:

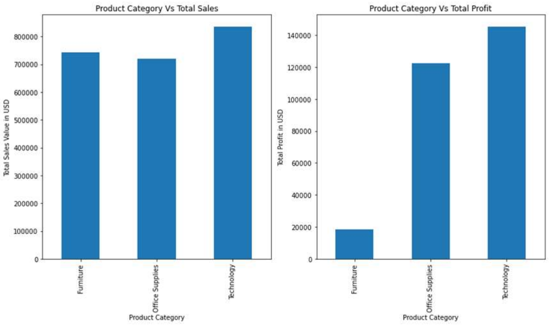


Fig: 7

Inference: In sales perspective all three categories are at the same level, but Furniture's profits are 5 to 6 folds lesser than the other two category. Further analysis on Sub-Category gives a better idea.

8. Analysis on the Sub-Category with Sales and profits:

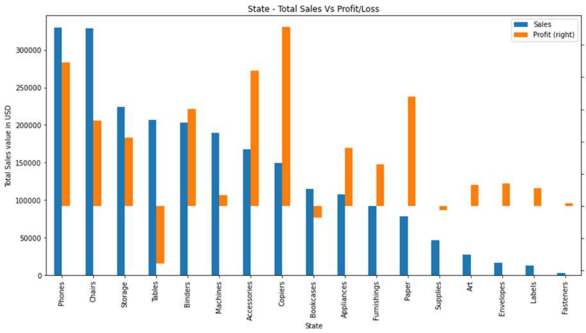


Fig: 8

Note: The profit is layered in the secondary Y axis.

Inference: Copiers, Accessories, Phones are the highly profited sub-category and all three belongs to Technology. Tables belongs to Furniture which is a huge loss among all the sub-categories. This supports the phenomena in category analysis.

9. Feature Selection for Modeling:

The first step in modelling is to decide on the dependant and independent variable for our model. The dependent variable here is Profit and for selecting the independent variable we have pre-processed the dataset in such a way that the variables which needs to be analysed are converted to numerical types. Now create a correlation matrix and plotted in seaborn to analyze the correlation between each and every attribute with each other.

The higher the values, correlation between those two variables are high. If it's a positive vale the relationship is directly proportional and if its negative then its inversely proportional.

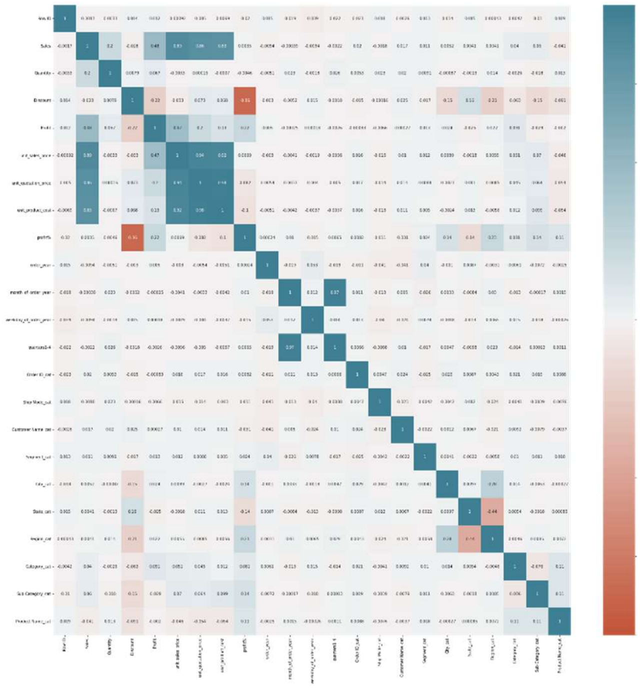


Fig: 9

From this we select the sales (0.48) and Discount (-0.22) attribute as the independent variable for modelling. The correlation between the independent variables has to checked if its correlation then drop that and go for the next value. In that aspect unit_sales_price (0.47), Profit% (0.22) and unit_quotation_price (0.20) is dropped.

C. *Construction of Model:*

Model 1: Simple Linear Regression

1. Build a simple linear regression model using the Scikit-learn package with dependent variable – Profit and Independent variable – Sales.
2. Quantify the variation by deriving R sqr values.
3. Predict the dependent variable (Profit).
4. Calculate the residuals – difference between the predicted data and actual data.
   Here is the scatter plot of Sales vs Profit, the red point are the actual data and the green points are the predictions from the model1.

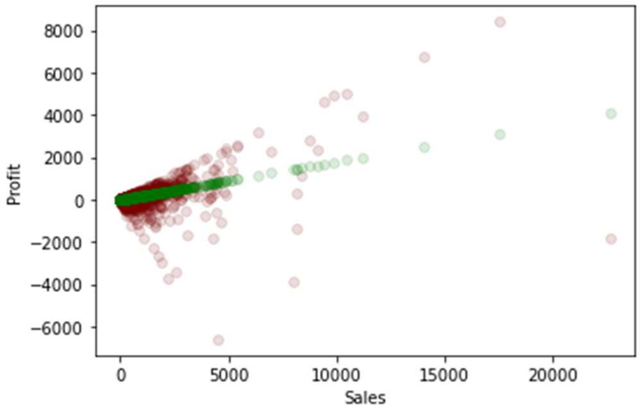


Fig: 10

Model 2: Multiple Linear Regression

Repeat the same procedure with 2 independent variables – Sales and Discount.

Model 1 explains only the relationship with Sales. Model 2 explains the influence of Discount variable added to it.

D. *Validation of Results:*

For a mix of continuous and categorical variables K-Fold cross validation can be done. Construct a K-fold model using the Scikit-learn package, specify the no. of splits to be done on the data. Run the model.

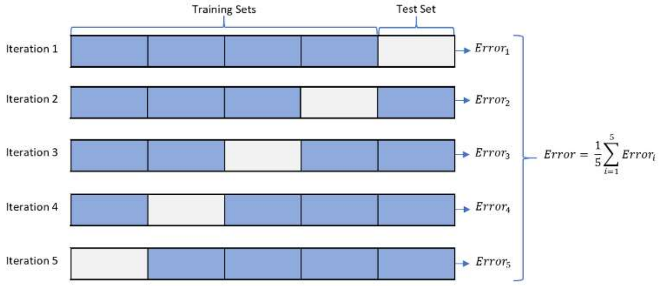


Fig: 11

Using the outputs of the models calculate the mean squared error values of all the iterations. Mean squared error value explains the variance of the prediction over the actual data and so it is good as it is low.

V. Findings, Reflections and Further Work

A. *Findings, Reflections:*

Question II B: Which customers contributes to the high business value and what are the products sold to them?

As mentioned in the point 5 and 6 of Analysis (Fig5 & 6), its is evident that customers from the west and east region contributes higher sales and even in the profits. Customers from California (West) leading the sales with almost 500 thousand USD which is 10 times higher than an average state sale and the profits are more than 70 thousand USD which is more than 10 times of the avg. state profit followed by New York (East) 2nd top of the table with almost same profits despite of three fourth of the sale of California. The products sold to these states are Binders, Papers, Accessories, Copiers in California, machines, phones and Binders in New York.

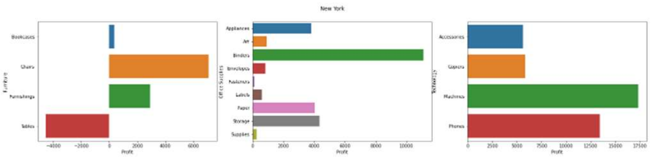


Fig: 12

Question II C: Which customers contributes to the loss and what are the products sold to them?

As mentioned in the point 5 and 6 of Analysis (Fig5 & 6), its is evident that customers from the central region contributes high loss but doing a significant business, Texas is holding 3rd position in the total sales highest loss of more than 20,000 USD whereas the average profit of a state is around 5,500 USD. This state consumes the other states profits and this is the biggest leak in the business where some action can be taken to improve the business profits. The products which is highly sold in loss to Texas is Binders, which is sold in a good profit will other states.

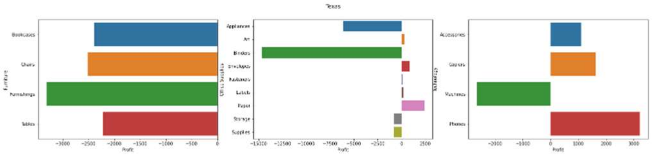
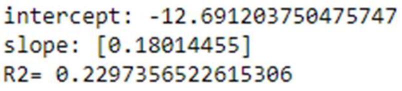
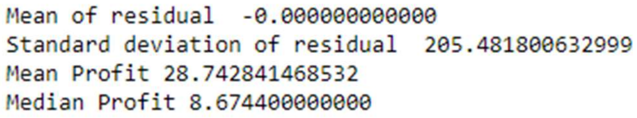


Fig: 13

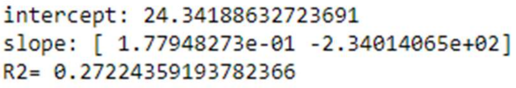
B. *Question II D:* Create a model to predict the profit of the order. Which are the attributes that contributes in deciding profits?

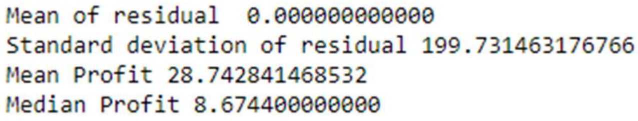
1. The results of Model 1: only Sales as independent variable.


```
intercept: -12.691203750475747
slope: [0.18014455]
R2= 0.2297356522615306
```


```
Mean of residual  -0.000000000000
Standard deviation of residual  205.481800632999
Mean Profit 28.742841468532
Median Profit 8.674400000000
```

2. The results of Model 2: Sales and Discount as independent variable.


```
intercept: 24.34188632723691
slope: [ 1.77948273e-01 -2.34014065e+02]
R2= 0.27224359193782366
```


```
Mean of residual  0.000000000000
Standard deviation of residual 199.731463176766
Mean Profit 28.742841468532
Median Profit 8.674400000000
```

Output parameters of Models:

- Intercept - is the fixed portion of price that's not affected by the independent variable.
- Slope - is the multiplier describing how an increase in one unit of independent variable increased the dependent variable.
- R squared – variations in the dependent variable due to independent variable.

3. Output of K-Fold cross validation:
   - Squared Error values of each iteration

K-Fold Iterations	Squared Error values
Iteration 1	60630.51
Iteration 2	73269.17
Iteration 3	19982.53
Iteration 4	22639.25
Iteration 5	28839.62
Mean Squared Error	41072.22

Findings:

- R1 value of the model 1 is 0.23 which says that 23% of the variation in profit can only be explained by Sales.
- R2 value of model 2 is 0.27 which is higher than the model 1 that evidences the addition of Discount variable explains more about the profit.
- The R2 values in both the models are very less because few attributes are only correlated to profit.
- Mean Squared Error values are very high which says that the variation between the prediction and the actual data is very high.
- Categorical attributes are not contributing much in this model, we have tried binning, transforming the data.

C. *Further Work:*

- In the descriptive analysis drilling down further on the states and sub-category, customer name and the product name can be figured out.
- In the modelling part, from this paper it is evident that Multiple Linear Regression explains well about the continuous data. There seems to be a limitation in exploring the categorical data. Need to try some other models to check the actual reason.

VI. REFERENCES

[1] Illustrating Statistical Procedures: Finding Meaning in Quantitative Data . 2020 May 15 : 61–139. Published online 2020 May 15.

[2] Procedia - Social and Behavioral Sciences, Volume 106, 10 December 2013, Pages 234-240

[3] Cross-Validation, January 2018, DOI:10.1016/B978-0-12-809633-8.20349-X, In book: Reference Module in Life Sciences Authors: Daniel Berrar, The Open University (UK)

[4] Pattern Recognition, Volume 69, September 2017, Pages 94-106, Error estimation based on variance analysis of k-fold cross-validation, Author links open overlay panelGaoxiaJiangWenjianWang

VII. WORD COUNTS

Total number of words in each section is mentioned as under,

1.	Introduction	- 201
2.	Research Questions	- 136
3.	Data (Materials)	- 242
4.	Analysis	- 992
5.	Findings, Reflections and Further Work	- 581