| Academic Year | Work Title | Assessment Number | Assessment Type |
|---|---|---|---|
| A21 | Beauty products analysis | A1 | Report |

*SIDDHARTHA SHRESHTA*
*2049848*
s.shrestha28@wlv.ac.uk

AAZAD KHADKA
*2051907*
A.Khadka7@wlv.ac.uk

# Table of Contents


I. Introduction ..... 1
II. Problem Statement ..... 1
III. Aim/objective of the work ..... 1
IV. Related works ..... 2
V. Methodology ..... 2
A. Data Collection: ..... 2
1) Data Source ..... 2
2) Data Description ..... 2
B. Understanding the features ..... 2
C. Analysing the data ..... 2
D. Data Cleaning and preprocessing : ..... 3
E. Data Visualization ..... 3
F. Results and findings : ..... 3
G. Block Diagram ..... 4
VI. Result and Discussion: ..... 4
f. Shifts and their sales record: ..... 6
g. Service sales in Different location: ..... 7
VII. Conclusion and Recommendation ..... 7
VIII. References ..... 7

***Abstract*—**

**Beauty industry is a huge industry with plenty of data to be stored and analyzed to expand the ongoing sales and stay up to date with the rising issues and deal with it swifter. The report below is a small example which uses the dataset of a beauty salon to provide meaningful insights related to product sales. The project utilizes the use of jupyter notebook and mongodb for storing and generating analysis of the dataset fed in these tools.**

## I. INTRODUCTION

Billions and billions of data are generated in the field of various types of businesses. Among these types, beauty industry is a huge one as suggested by the article [1] which estimates the beauty industry spending amount to be at least 483 billion dollars. Such a huge industry-based companies need to depend on big data to visualize and analyze critical information such as profit or loss, product sales (location wise), product restock and so on. Various insight generating tools are used to generate the analysis. Using these tools, a business entity can capitalize and improve its current business situation to expand even further in this never so stopping highly competitive industry

## II. PROBLEM STATEMENT

As per [2]the phrase Big Data refers to the collection of a large quantity of data from numerous sources at a high volume and velocity using a range of digital devices. As big data expands at a rapid rate, it becomes increasingly difficult to handle, manage, and analyze using standard technologies. The procedure for examining a huge number of data sets. In order to do that a business company has to spend large sum of its revenue and time as well as resources. Using the traditional way of data analysis. In order solve that issue this works focuses on the modern statistical and qualitative as well as quantitative data analysis approach will be initiated in this work.

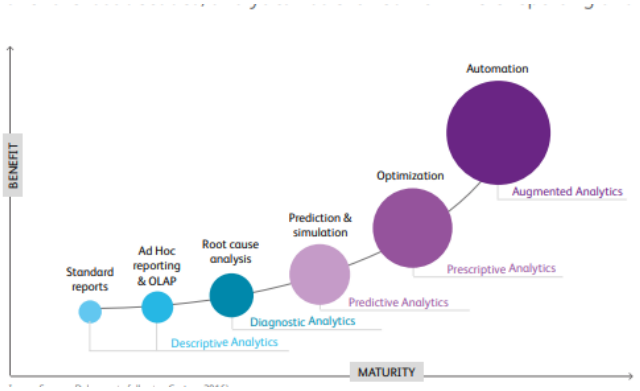


*Figure 1 Beneficial Results of the Data analysis and Business Intelligence*

## III. AIM/OBJECTIVE OF THE WORK

The aim of this analysis report is as follow:

- To generate a concrete feature analysis
- To know the Possible loop holes that affected business for long term

Some of the main objectives of this analytical report are as follows:

- *To understand and use the data analysis principles.*
- *To be able to work with different big data processing methods.*
- *To be able to use data analysis tools and technologies as advantage*.
- To be able to generate multiple insights on the given data.
- To be able to make logical decisions and plans for the future of the company.

- To generate experimental analysis of new strategy.

## IV. Related works

This article by [3] examines the key aspects of refined sales management in the context of big data.
The main points of sales management include building a management system for sales and bettering the sales management information system, evaluation management system, and strengthening the flow of inside sales. The author investigates the data storage place creation, searching and purifying the data, creating of models for predicting, and result analysis allocation by combining the key points of data analysis under big data. The goal of this article is to assist people in fully utilizing the benefits of big data technology applications while also promoting the lucrative development of the company.

In today's world data is generate by every aspect of the world either it's a simple shop or it's a multi billions company. This paper [4] proposed a system developed from java using WebCrawler and analyze the sales data of automobile company. Thus, the analysis will degrade the amount of waste materials and utilize the firms and materials which are essential and stocked.
This article [5] analyzed the electricity sales data which is affected by various factors. RNN model was developed with the electricity sales big data and each feature were analyzed. Multiple neural network models were built and trained on multiple time spans and prediction were done. Experiments on a power supply company's historical electricity sales data set reveal that this technology outperforms standard machine learning methods in terms of accuracy and efficiency.
Restaurant operators must precisely anticipate the number of potential guests in order to run an efficient and profitable business. The research [6],a model was developed to predict peoples count to their restaurant in future. Regression model was made Using KNN, Random Forest and XGBoost.

Big data has progressively become a significant aspect of people's lives in recent years. The software that was developed in those time period provided people to give reviews, opinions, their experience etc. This type of review data enables firms to do service defect analysis. By the data gathered from each people any business can analyze it and find their flaws and move themselves towards the betterment. This study [7] proposed to build a analysis model that utilized reviews provided by user and creating a correlation between different parameters and features of the reviews and then creating a marketing plan accordingly as per the analysis report generated.

According to the findings of the research [8]. Price seems to be the most dependent aspect for the sales of any products however the quality of products may counter that thought. The effects of many variables on the sales of famous science publications vary according to the number of total sales by other customer

## V. Methodology

The business data is of beauty product and service provider of Nepal. The data has all the necessary sales data to be analyzed. In order to process the data and generate a concrete insightful report many queries and statistical report was generated using different big data analysis tools like mongo DB, Python and elastic search.

### *A. Data Collection:*

#### *1) Data Source*

The data set was acquired from the business company itself. The dataset which was acquired are structured in a way that it stores most of the sales related data. The data is highly leveraged on the product and service details.

#### *2) Data Description*

Looking at the features of the dataset there **24 columns** of features**, and 34822 rows** data of those features. It contains data of sales order, sales id, real product price, product selling price, product name, billing quantity, billing date m category name, sub-category name, product type, billing time, discount, branch code, shift name, void value.

### *B. Understanding the features*

In order to understand each column data into more depth and to make it easier to process firstly the data types are checked and found that the dataset contains, null values, integer values, string data, data/time. Keeping the data types in reference an initial hypothesis was created and on the basis of that proper relationship was developed and bar, graph, heatmap etc. we're being used.

### *C. Analysing the data*

For data analysis part technologies like mongo DB, python and elastic search was used. Python comes with libraries like pandas which was implemented in order to calculate the some of the logical operations on the numerical field and was also used to generate the calculation report like profit and loss, min-max, standard deviation, normalization of numerical data etc. For the visualization of data **matplotlib** and **seaborn** plot was used to create pie-chart and bar plot of multiple fields

.

*D. Data Cleaning and preprocessing :*

Some of the columns contained null values in the dataset, in-order to solve that issue mean was calculated on that column and those null fields were replaced with the mean value. All those irrelevant columns like product-id, sales-id etc were dropped from the dataset. Some of the categorical values were one hot encoded and were added to separate

*E. Data Visualization*

For data visualization, matplotlib and seaborn was used. The relation between the features and the target was visualized and observed through different bars, charts, heatmaps, and histograms. These visualization techniques provide a general overview over the dataset and helps to understand the general relation between the different features within the dataset.

*F. Results and findings :*

After the preprocessing of the data some queries were integrated between multiple fields and generated insight and plotted visual representation like tabular representation, pie chart, bar-plot etc.

Tabular reports were used for those mass number of logical and numerical operation which provided clear insights with both numerical and string data which are inter related.

*G. Block Diagram*

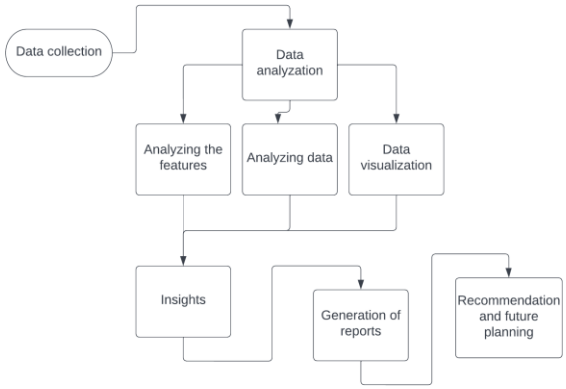


*Figure 2 Flow/Block diagram of Methodology*

## VI. Result and Discussion:

### a. Most sold product type:

The organization provides two types of services i.e., “Product” and “Services” among them. Service was most sold. The total sales count of service was 28607 which is 82.15% of the total sales. That means the company is generating most of their revenue from the service they provide

```
> db.Beauty.find({product_type: "Product"}).count()
< 6215
> db.Beauty.find({product_type: "Service"}).count()
< 28607
```

```
In [48]: plt.figure(figsize=(10,10))
         plt.title("Product Type")
         df["product_type"].value_counts().plot(kind="pie",autopct='%1.2f%%')
Out[48]: <AxesSubplot:title={'center':'Product Type'}, ylabel='product_type'>
```

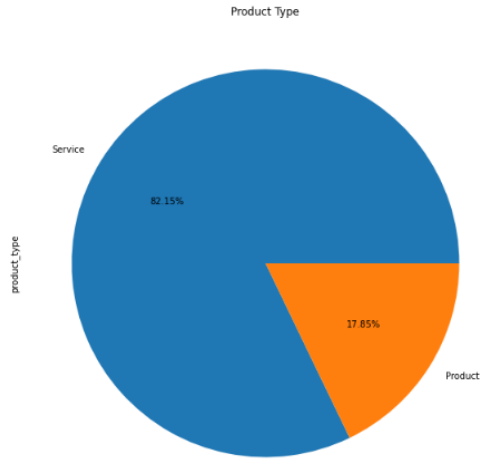


*Figure 3 Pie chart for type of business sells*

b. Most sold product:

The most sold product was Gents-Cut and Style Haircut which is one of their service for men haircut. The total sales count for the service is **28372** which is **82.15%** of total sells and other services who had sales higher that 2000 are:

- GENTS-CUT AND STYLE HAIRCUT :28372
- LADIES-CUT AND STYLE HAIRCUT (TOP STYLIST): 7622
- LADIES-THREADING-EYE BROW: 6348
- LADIES-THREADING-UPPER-LIP: 3154
- GENTS-MASSAGE-HEAD: 2986
- GENTS-COLOUR-DARK BROWN: 2541
- LADIES-MASSAGE-HEAD :2120
- GENTS-CUT AND STYLE-STYLE SHAVING: 2113
- GENTS-CUT AND STYLE-BEARD TRIMING: 2064

```
In [17]: top_sales=most_sales[most_sales > 2000]
         print(top_sales.to_markdown())
```

| | product_name |
|---|---|
| GENTS-CUT AND STYLE HAIRCUT | 28372 |
| LADIES-CUT AND STYLE HAIRCUT(TOP STYLIST) | 7622 |
| LADIES-THREADING-EYE BROW | 6348 |
| LADIES-THREADING-UPPER-LIP | 3154 |
| GENTS-MASSAGE-HEAD | 2986 |
| GENTS-COLOUR-DARK BROWN | 2541 |
| LADIES-MASSAGE-HEAD | 2120 |
| GENTS-CUT AND STYLE-STYLE SHAVING | 2113 |
| GENTS-CUT AND STYLE-BEARD TRIMING | 2064 |

*Figure 6 Most sold products table*

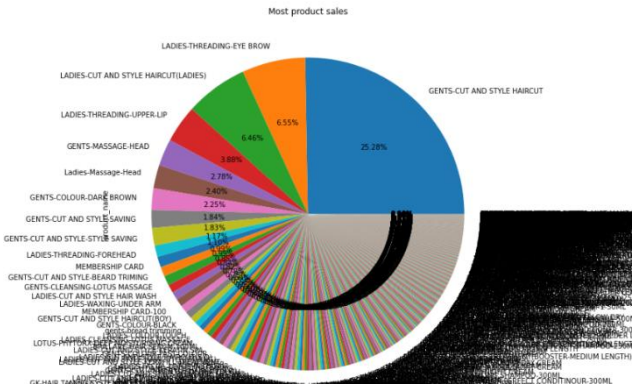


*Figure 5 Most sold product pie chart*

c. Lowest sales:

There were 730 items which were lowest in sales. some of them are show in figure below.

```
low_sales=lowest_sales[lowest_sales<100]
low=lowest_sales[lowest_sales<100].count()
print(low_sales.to_markdown())
print(low)
```

| | |
|---|---|
| GENTS-FACIAL-LOTUS DERMO SPA WHITENING AND LIGHTENING | 10 |
| L-Colour-global (long length) | 10 |
| CHAMAL | 10 |
| LADIES-COLOUR HIGHLIGHT TWO COLOUR | 10 |
| LOTUS-PHYTORX-DEEP PORE CLEANSING FACEWASH | 10 |
| GENTS-FACIAL-LOTUS PRV PAPAYA | 10 |
| STREAX-CANVO LINE STRAIGHTNING SERUM-100ML | 10 |
| ladies-colour-global (medium to long ) | 10 |
| LOTUS-PHYTORX-SPF-50 MATTE | 10 |
| LADIES-IRON STRAIGHT (TEMPORARY) | 10 |
| gents-colour-fast | 10 |
| GENTS-CUT AND STYLE-KERATIN SHOULDER | 10 |
| GENTS-FACIAL-LOTUS PRV KIWI FRUIT | 10 |
| BRILLARE-STYLE CARE CONDITIONER-125ML | 11 |
| COLOUR-PERFECT-22/0 | 11 |
| LOREAL-HOMME MAT SCULPTING POMADE-4-80ML | 11 |
| GENTS-WAX-UNDER ARM | 11 |
| COLOUR-SCHWARZKOPF-7-55 | 11 |
| LOTUS-PHYTORX-ANTI AGEING SERUM(INTENSIVE REPAIR) | 11 |

*Figure 4 least sold product table*

d. Profit and loss:

The company didn't go under loss because they had many null values on loss field.

| | product_name | real_product_price | selling_product_price | Profit | Loss |
|---|---|---|---|---|---|
| 16 | GENTS-CUT AND STYLE-THREADING | 150.0 | 200.0 | 50.0 | None |
| 80 | TISSUE PAPER | 125.0 | 130.0 | 5.0 | None |
| 83 | LADIES-STRAIGHT-MEDIUM LeANTH | 5000.0 | 5500.0 | 500.0 | None |
| 120 | TISSUE PAPER | 125.0 | 130.0 | 5.0 | None |
| 142 | TISSUE PAPER | 125.0 | 130.0 | 5.0 | None |
| ... | ... | ... | ... | ... | ... |
| 102112 | LOTUS-PHYTORX-JAPANESE SAKURA FACE WASH | 710.0 | 760.0 | 50.0 | None |
| 102161 | LADIES-CUT AND STYLE-TREATMENT(BOOSTER-SHOLDER... | 1150.0 | 1200.0 | 50.0 | None |
| 102168 | MEMBERSHIP CARD-100 | 100.0 | 250.0 | 150.0 | None |
| 102173 | EYE LYSIS | 200.0 | 300.0 | 100.0 | None |
| 102210 | TIPS | 100.0 | 235.0 | 135.0 | None |

*Figure 7 Profit and loss table*

```
product_name              0
real_product_price        0
selling_product_price     0
Profit                    0
Loss                   2269
```

e. Categories insights:

They have **16 categories** of them of their products and services. they are as follow:

- ASTABERRY
- BRILLARE
- GATSBY
- GK
- LIFE & CARE'
- LOREAL',
- LOTUS'
- PRODUCT
- RICHFEEL
- SERVICE
- SCHWARZKOPF
- SILK HAIR
- STREAX
- WELLA
- gents
- ladies

```
> db.Beauty.distinct("category_name")
< [
    'ASTABERRY',    'BRILLARE',
    'GATSBY',       'GK',
    'LIFE & CARE',  'LOREAL',
    'LOTUS',        'PRODUCT',
    'RICHFEEL',     'SARVICE',
    'SCHWARZKOPF',  'SILK HAIR',
    'STREAX',       'WELLA',
    'gents',        'ladies'
  ]
```

***Among those 16 products "SCHWARZKOPF" was most sold.***

```
> db.Beauty.find({category_name: "ASTABERRY"}).count()
< 29
> db.Beauty.find({category_name: "BRILLARE"}).count()
< 117
> db.Beauty.find({category_name: "GATSBY"}).count()
< 9
> db.Beauty.find({category_name: "LIFE & CARE"}).count()
< 295
> db.Beauty.find({category_name: "LOTUS"}).count()
< 66
> db.Beauty.find({category_name: "RICHFEEL"}).count()
> db.Beauty.find({category_name: "PRODUCT"}).count()
< 27619
> db.Beauty.find({category_name: "SARVICE"}).count()
< 5232
> db.Beauty.find({category_name: "SILK HAIR"}).count()
< 43
> db.Beauty.find({category_name: "WELLA"}).count()
< 347
> db.Beauty.find({category_name: "ladies"}).count()
< 13
e: "SCHWARZKOPF"}).count()
e: "STREAX"}).count()
e: "gents"}).count()
e: "BRILLARE"}).count()
e: "GK"}).count()
```

*f. Shifts and their sales record:*

After running the query, we found that there were 3 branches of the business outlet they are:

- Koteshwor
- Maharajgunj
- Pulchowk.

The total sales of the business products and services at different location was 8153 products and services at Koteshwor branch, **15079** at Maharajgunj branch and 11590 and services at **Pulchowk** branch.

```
> db.Beauty.find({"branch_name":"Pulchowk"}).count()
< 11590
> db.Beauty.distinct("branch_name")
< [ 'Koteshwor', 'Maharajgunj', 'Pulchowk' ]
> db.Beauty.find({"branch_name":"Pulchowk"}).count()
< 11590
> db.Beauty.find({"branch_name":"Maharajgunj"}).count()
< 15079
> db.Beauty.find({"branch_name":"Koteshwor"}).count()
< 8153
```

*Figure 8 Branch sales count*

By which we can conclude that there had been **highest sales on Maharajgunj Branch** in total sales withing the dataset provided to us.
The company had 3 shifts for their business conduct time; Morning, Evening and Afternoon.
Looking at the analysis report there were **5699 total sales during** morning time and **9298 at evening** and **19717 at**

**afternoon** shift which means the company had **most of its sales at afternoon time.**

```
> db.Beauty.distinct("shift_name")
< [ 'Afternoon', 'Evening', 'Morning' ]
```

```
> db.Beauty.distinct("shift_name")
< [ 'Afternoon', 'Evening', 'Morning' ]
> db.Beauty.find({shift_name: "Morning"}).count()
< 5699
> db.Beauty.find({shift_name: "Evening"}).count()
< 9298
> db.Beauty.find({shift_name: "Afternoon"}).count()
< 19717
```

*Figure 9 Shift count*

*g. Service sales in Different location:*

The above figure shows that there were **9645** service sales in

```
> db.Beauty.find({product_type: "Service",branch_name: "Pulchowk"}).count()
< 9645
> db.Beauty.find({product_type: "Service",branch_name: "Koteshwor"}).count()
< 6209
> db.Beauty.find({product_type: "Service",branch_name: "Maharajgunj"}).count()
< 12753
```

*Figure 10 service in different location*

pulchowk , **6209** in **koteshowr** branch and **12753** services in **Mharajgunj** branch . Highest was on Maharajgunj Branch

h. Product sales in Different location:

```
> db.Beauty.find({product_type: "Product",branch_name: "Maharajgunj"}).count()
< 2326
> db.Beauty.find({product_type: "Product",branch_name: "Koteswhor"}).count()
< 0
> db.Beauty.find({product_type: "Product",branch_name: "Pulchowk"}).count()
< 1945
```

*Figure 11 Product sales in different locations*

In the above figure it showed that none of the product were sold in koteshwor branch

## VII. Conclusion and Recommendation

By the above finding the business should remove selling products category to koteshowr branch because the product category sells were 0 and the Gent hait style service should be enchanced and more of that particular service should be provided. The organization should focus on expanding services on their maharajgunj branch since most of its revenew is generated from that branch. The business sells revenue is generated most by their services type which is 82.15% of their total sells revenue.

In this way this work was able to use the modern analytical techniques and tools to generate a concrete insight on a real-world beauty service data of a company and was able to recommend some of the future plannings for the organizational growth.

## VIII. References


[1] J. Sickler, "terakeet," 15 July 2021. [Online]. Available: https://terakeet.com/blog/beauty-industry/. [Accessed 2 May 2022].

[2] **Wani** and S. Jabin, "Big Data: Issues, Challenges, and Techniques in Business Intelligence," in *Big Data Analytics*, 2018, pp. 613-628}.

[3] W. Shan, 2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), Taiyuan: IEEE, 2020.

[4] Q. Zhang, H. Zhan and J. Yu, "Car Sales Analysis Based on the Application of Big Data," *International Congress of Information and Communication Technology (ICICT 2017)*, vol. 107, pp. 436-441, 2017.

[5] C. Qi, S. Mingyue, M. Na, W. Honggang, J. Yanhong and Z. Jing, Regional Electricity Sales Forecasting Research Based on Big Data Application Service Platform, Xi'An: IEEE, 2020.

[6] X. MA, Y. TIAN, C. LUO and Y. ZHANG, Predicting Future Visitors Of Restaurants Using Big Data, Chengdu: IEEE, 2018.

[7] S. Liu, How to Realize Precision Marketing in Catering Industry through Big Data, Chengdu: IEEE, 2021.

[8] Q. Zhong and Y. Wang, Analysis of Factors Affecting the Sales of Popular Science Books Based on Big Data, Zhengzhou: IEEE, 2020.