A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Alexandra Munik Silva |
| *Student Number* | 2024268 |
| *Module Title* | Strategic Thinking |
| *Assessment Title* | CA2 Strategic Thinking |
| *Assessment Due Date* | Sunday, 15 December 2024 |
| *Date of Submission* | Sunday, 5 December 2024 |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Contents

[Strategic overview of the business problem 1](#_Toc184327014)

[Project plan 1](#_Toc184327015)

[Business understanding 2](#_Toc184327016)

[Data understanding 2](#_Toc184327017)

[Data preparation 2](#_Toc184327018)

[Data Cleaning 3](#_Toc184327019)

[Outliers 3](#_Toc184327020)

[Statistic 4](#_Toc184327021)

[Insights 5](#_Toc184327022)

[Machine Learning 6](#_Toc184327023)

[Feature Importances 6](#_Toc184327024)

[Predictive Inference 6](#_Toc184327025)

[Insigth 7](#_Toc184327026)

[Heatmaps 7](#_Toc184327027)

[Insights 8](#_Toc184327029)

[Conclusion 9](#_Toc184327030)

[References 9](#_Toc184327031)

# Introduction

Data analysis is a weapon for firms not only willing to know their past but to predict and shape their future.

In the keenly competitive world of e-commerce, decisions that are data-driven is the recipe for success in outdoing the rivals.

This paper looks at the possible ways an e-commerce firm can make use of data analysis to have an impact on its sales strategies, enrich customer experience, and in the end drive profitability.

The objective is to determine which the best-selling products are, how reviews relate to sales, and the effect of price on purchasing behavior.

The main point of this work is that without data analysis, companies might be making decisions under the blindness of operating on assumptions or inadequate information, which can eventually trickle down to flawed strategic choices.

As we shall see, quantitative analysis not only improves marketing campaigns and inventory management; it also brightens up invaluable insights that could help shape the future of businesses.

With the application of machine learning methods and statistical visualizations, it became possible to predict the future performance of new products and lost the key variables affecting sales most such as product reviews.

This study proves that with the use of data, tically an company can make decisions based on information that will enhance its market position and profitability.

# Strategic overview of the business problem

The problem presented in the first part of the project was that although data analysis tools are extremely effective for the success of a good business strategy, some companies do not use data analysis as a competitive advantage. In this part of the project, it will be presented how a company could apply data analysis to its strategy so that this tool brings agility and more assertiveness in decision making.

“On the other hand, perhaps your operational business processes aren't much

different from anybody else's, but you feel you compete on making the best

decisions. Maybe you can choose the best locations for your stores-if so, you're

probably doing it analytically. You may build scale through mergers and

acquisitions, and select only the best candidates for such combinations. Most don't

work out well, according to widely publicized research, but yours do. If so, you're

probably not making those decisions primarily on intuition. Good decisions usually

have systematically assembled data and analysis behind them.” (Davenport and Harris, 2017)

In the book (Davenport and Harris, 2017) he mentions how the data analysis tool can help in a company's decision-making and can become a competitive advantage, placing the company ahead of other companies in the same sector that decide not to use data analysis. data in their decisions.

# Project plan

At this stage, a dataset was collected with information about sales made by an e-commerce. The initial idea is to understand the best-selling products so that more accurate predictions can be made.

In this case, is a e-commerce that sells different types of products, such as household appliances, toys, clothes, etc. The intention is to understand this company's sales to plan inventory, explore new suppliers, understand the public for future offers, pricing, among other points, always aiming for the company's profitability and assertiveness. As a result, we can prove the idea that good data analysis can be extremely important for a company's future decisions.

“The wide adoption of customer relationship management, or CRM, and supply chain management software has allowed enterprises to fully interface and integrate their demand and supply chains. Based on this integration, they are better able to capture up-tothe-minute data about demand for a particular product, as well as data of similar granularity about the supply of corresponding data. Analysing these two data streams, organizations optimize the price of a particular product along several dimensions so demand meets available supply; for example, the price of a product may be different through one channel (such as the Web) than through another (such as a retail store). Price optimization allows any type of organization to maximize profit margins for each item sold while reducing inventory.” ((Kohavi, Rothleder and Simoudis, 2002)

# 

# Business understanding

An online company that sells various types of products was selected, and in the dataset in question we have sales information for these products over a period of 12 months, also containing important variables such as price and evaluation of these products during this period.

The challenge of this dataframe was to understand what the best-selling product would be in the year in question, whether the reviews influenced the increase in sales of a specific item or whether the price impacted sales. The data frame also shows us important variables about quantities sold according to specific months. In this case we will explore the importance of each variable with the intention of absorbing important insights for the company model studied in question.

# Data understanding

Basically, the data frame describes the ID of the products sold, product name, category, price, product reviews and units sold per month.

This data frame is a file with 1000 rows and 18 columns as described below:

* Product id: numbers that identify products
* Product name: name the products (numbers)
* Category: This case we have 7 categories which is Books, toys, Home & Kitchen, clothing, Health, Sport, electronics and clothing.
* Price: Price Review score: Review provided from customers
* Review count: number of reviews
* Sales month X: Sales month 1 until Month 12

Data preparation and Data Cleaning

For Data preparation, the necessary libraries for reading, interpreting, plotting data, recognizing the data type of the data frame and plotting graphs were imported.

The libraries used were: Pandas, seaborn, numpy, sklearn.decomposition, matplotlib.pyplot, seaborn. To understand the data frame, the following formulas were used with their following interpretations:

* pd.read\_csv: Used to read the data frame.
* df.head: Used to show the first 5 rows of the dataframe for better visualization.
* print(df.columns): Used to print the column names.
* df.shape: Used to understand the size of the Data frame in this case consisting of 1000 rows and 18 columns.
* df.dtypes: Shows the data types contained in the Dataframe, file size, number of rows and columns.

After analyzing the data frame, the cleaning process began with the following functions:

* df.drop: Function used to discard some columns, in this case the “product\_name” and “product\_id” columns were discarded because according to the proposed analysis, they would not make a big difference in the data frame as they only deal with product names and ids.
* df.isnull: Function for identifying missing values, in this case we have the result of “false” indicating that there are no missing values ​​in the data frame.

# Outliers

An Outlier is a data item/object that deviates significantly from the rest of the (so-called normal) objects. Identifying outliers is important in statistics and data analysis because they can have a significant impact on the results of statistical analyses. The analysis for outlier detection is referred to as outlier mining.  (GeeksforGeeks, 2021)

To begin the process of applying Statistical and Machine Learning data, techniques were applied to understand outliers and remove them so that no discrepancies occur in the results. For this, the following functions extracted from the scipy.stats library were applied. IQR and Z-scores techniques were used to identify outliers.

* IQR: It helps identify outliers and variability in data.
* Z-score: Identify how many standard deviation a data point is from the mean.

After identification we had the following result:

Tabela

Descrição gerada automaticamente

After analysis, outliers were removed so that the statistical analysis could be optimized.

# Statistic

Using the formulas df.groupby, category\_sales.idxmax(), category\_sales.max() , category\_sales\_sorted.plot we can view the following graph and obtain important information for analysis.

* df.groupby: Groups the data in a DataFrame based on one or more columns, allowing aggregate calculations into groups.
* category\_sales.idxmax(): Returns the index of the maximum value in a series or group
* category\_sales.max(): Returns the maximum value in a series or group
* category\_sales\_sorted.plot: Generates a graph from an ordered series or DataFrame, visualizing data in graphical format.

Gráfico, Gráfico de barras

Descrição gerada automaticamente  
In the graph drawn up above, we obtain information that the most sold products by category in the year were books, toys and sporting goods respectively, while the least sold were Home & Kitchen, clothing and Health, however it is important to emphasize that even the least sold products have a very large weight in sales, since we are talking about millions of products sold per category per year.

# Insights

* -Given that the company sells more books, it might be interesting to implement greater marketing on this specific product, focusing on promotional campaigns to attract new consumers.
* -Given that the least sold category was Home and Kitchen, it would be interesting to first understand the weakness of this category, in this case it could be understood more about the competitors, in relation to price and variety of products for example.
* For some categories such as Sports, Electronics and Clothing perhaps a good technique to apply would be ways of differentiating these products, through prices or even applying Cross Selling techniques connecting some products between them (for example, the sale of books related to sport with sporting goods).

# Machine Learning

For machine learning, some techniques were applied to understand and make predictions, tools such as Feature Importances, Predictive Inference and Heatmaps were used to make predictions and understand better the data.

# Feature Importances

The model.feature\_importances\_ formula returns the relative importance of each feature in a tree-based machine learning model, such as Random Forest or Gradient Boosting. It helps identify which variables have the greatest influence on the model’s predictions.

Feature importance refers to a class of techniques for assigning scores to input features to a predictive model that indicates the relative importance of each feature when making a prediction. (Brownlee, 2020)

Texto

Descrição gerada automaticamente

In this case we had greater relevance in the features on the number of products sold per month, in which we can see that they present results above 0.1 indicating the importance of these values in the dataframe.

# Predictive Inference

Predictive Inference is the process of using a statistical or machine learning model to predict future or unknown values based on available data. The focus is on accurately estimating responses to new cases, regardless of understanding the underlying mechanisms. (Sanders, 2019)

After that, another tool was applied to predict what the introduction of a new product would be like in the company. We then had the following result through the Predictive Inference technique applied below:

Texto

Descrição gerada automaticamente

We can predict that a new product would have an acceptable performance in the first 6 months, as it would represent 37% of sales compared to the current sales of the best-selling category (books).

# Insigth

* The introduction of a new product would be a good strategy for the company; however, it would be interesting in this case to consider the best-selling products and relate this new product to the others so that the company has more possibilities to apply strategies based on the best-selling products (upselling).

# Heatmaps

“The visual nature of heatmaps allows for immediate recognition of patterns, such as clusters, trends, and anomalies. This makes heatmaps an effective tool for exploratory data analysis” (Bothma, 2024)

This graph shows how the variables are related to each other, but in this case there are not many visible correlations, values ​​1 correspond to perfect correlation (when one variable increases, another also increases) -1 negative correlation (when there is an increase in one variable and a decrease in other) 0 there is no correlation.

Gráfico, Gráfico de mapa de árvore

Descrição gerada automaticamente

# We can notice that most of the correlations are weak, close to 0, we can also verify in this case that the review variable has a positive correlation in relation to the number of sales. In other words, the higher the review score, the higher the units sold. Which proves our initial idea that reviews impact sales, in this case we can gain some insight into this.

# Insights

* This company in question could invest in good marketing or simply adopt post-sales research as a strategy to increase the number of reviews for this e-commerce.
* Another idea would be to understand reviews through other analyses to understand and guarantee customer satisfaction since through the graph we can see that the higher the Review score, the greater the number of sales.

# Conclusion

After the analyzes carried out above, we can conclude that data analysis is a very important and relevant tool for a company, without this tool, a company would be left to make assumptions, considering poor information without much relevance for the company.

Many companies underestimate the power of data analysis, often leading the company to bankruptcy due to lack of information and often due to applying bias in business strategies.

Through this data frame we can extract several insights such as:

* Improvements in marketing campaigns more focused on specific products. Pay attention to customers in some specific categories with the intention of understanding the products they are looking for and their level of satisfaction with the product.
* Pay attention to competitors, understand prices, product variety, etc.
* Apply Upselling techniques with the intention of increasing sales of some products through products that have strong selling power in the company.
* We were able to predict how a new product would perform in future years.
* We were also able to verify the impact between variables, that is, the higher the Review Score of a given category, the greater the number of products sold.

In this case, we have an interesting observation about this analysis, if we were to apply Bias in this analysis, we would automatically go to the price factor, but in this case the price factor was not as relevant according to the analysis. Different from the Review factor, which can give us more relevant information and consequently more accurate insights, information about sales, we can have access to the behavior of products sold, customers, employees and the company within a general scenario.

Thus proving the problem initially addressed, about companies that do not invest in data analysis, this project aimed to show how easy it is to make decisions based on data and facts as shown above, with the intention of minimizing expenses, opening up possibilities for new strategies, new products and new perspectives in relation to the market and customers.

# References

Bothma, J. (2024). *Seaborn Heatmaps: A Guide to Data Visualization*. [online] Datacamp.com. Available at: https://www.datacamp.com/tutorial/seaborn-heatmaps.

Brownlee, J. (2020). *How to Calculate Feature Importance With Python*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/calculate-feature-importance-with-python/.

Davenport, T.H. and Harris, J.G. (2017). *Competing on analytics : the new science of winning*. Boston, Massachusetts: Harvard Business Review Press.

GeeksforGeeks (2021). *Detect and Remove the Outliers using Python*. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/detect-and-remove-the-outliers-using-python/.

GeeksforGeeks. (2018). *Exploratory Data Analysis in Python*. [online] Available at: https://www.geeksforgeeks.org/exploratory-data-analysis-in-python/.

Kohavi, R., Rothleder, N.J. and Simoudis, E. (2002). Emerging trends in business analytics. *Communications of the ACM*, 45(8). doi:https://doi.org/10.1145/545151.545177.

Sanders, N. (2019). A Balanced Perspective on Prediction and Inference for Data Science in Industry. *Issue 1*. doi:https://doi.org/10.1162/99608f92.644ef4a4.

Versloot, C. (2022). *Machine learning articles*. [online] GitHub. Available at: https://github.com/christianversloot/machine-learning-articles/blob/main/how-to-predict-new-samples-with-your-keras-model.md.