Python Driven EDA & Data Visualization in Retail

Muhammad Riad Hossain
AMOD – Big Data Analytics
Trent University
Peterborough, Ontario, Canada
muhammadhossain@trentu.ca

Muhammad Sajid Salman
AMOD – Big Data Analytics
Trent University
Peterborough, Ontario, Canada
muhammadsajidsalman@trentu.ca

***Abstract*— This comprehensive study employs a robust retail dataset from Kaggle to conduct an exploratory data analysis (EDA), to extract insights into consumer behavior, sales performance, and operational effectiveness. Through rigorous exploratory data analysis (EDA), it highlights the intricacies of sales distributions, profit margins, and purchase volumes, pinpointing the significant role of outliers. Bivariate and multivariate analyses reveal the impact of geographical and segment-based factors on sales, alongside the evolution of profit trends over time.**

Statistical testing forms the core of the research, with regression analysis disclosing a negative correlation between discounts and sales, indicating the need for a refined discounting strategy. ANOVA testing illustrates the critical influence of product categories on sales outcomes, overshadowing the less pronounced effects of segments and regions. Additionally, a Chi-Square Independence Test confirms the non-association between shipping methods and regions, suggesting opportunities for shipping optimization.

The study's findings empower retail managers to make informed decisions on product category management and discount policies. Furthermore, the lack of dependency on regions for shipping preferences may allow for streamlined logistical operations. This comprehensive data-driven approach provides a valuable benchmark for retail industry practices, aiming to bolster customer satisfaction and enhance sales efficacy within the competitive retail market.

Keywords*— *Exploratory Data Analysis (EDA), Data Visualization, Statistical Testing, Retail, Sales, Customer Behavior.

I. Introduction

The retail industry, marked by its dynamic and competitive nature, constantly evolves due to consumer behavior, market trends, and operational strategies. Our study utilizes a Kaggle dataset, exploring diverse aspects of retail sales to gain nuanced insights into transactional patterns, customer interactions, and regional market dynamics. Through data-driven methodologies, we aim to contribute valuable knowledge for strategic decision-making in retail enterprises.

Our primary goal is to conduct a thorough exploratory data analysis of the retail dataset, revealing trends and patterns across various dimensions. We seek to understand consumer behavior, assess sales performance, and evaluate operational efficiency. Utilizing statistical testing, we investigate the impact of discounts, variations in product categories, and the association between shipping modes and regions. Our aim is to provide actionable insights empowering retail decision-makers to refine strategies, optimize product categories, and streamline operations.

This study covers diverse dimensions within the retail sector, including order and customer information, geographic data, sales categories, and product metrics. The significance lies in the dataset's comprehensiveness, enabling an in-depth analysis beyond surface-level observations. The findings not only apply to the specific dataset but also serve as a broader reference for the retail industry. Decision-makers can use these insights to enhance customer experiences, tailor marketing strategies, and improve overall business performance.

II. Previous Work

Previous research in the retail domain has extensively explored various aspects of consumer behavior, market trends, and operational efficiency. Studies have delved into the impact of discounts on sales, the significance of product categories, and the role of geographic factors in shaping market dynamics. Existing literature also highlights the importance of data-driven decision-making in the retail sector. However, despite the wealth of research, there is a noticeable gap in the integration of comprehensive datasets for a holistic analysis that spans multiple dimensions simultaneously.

Previous investigations have delved into the prediction of sales within the retail sector by leveraging historical data. In 2015, Harsoor and Patil (Harsoor & Patil, 2015) undertook a study focused on forecasting the sales of Walmart Stores. Their approach incorporated big data applications, specifically Hadoop, MapReduce, and Hive, to ensure the efficient management of resources.

A data scientist named Michael Crown (Crown, 2016) engaged in a similar dataset analysis, centering his efforts on time series forecasting and non-seasonal ARIMA models. Crown utilized ARIMA modeling techniques to generate one year of weekly forecasts, drawing from 2.75 years of sales data. The dataset encompassed various features such as store details, department information, dates, weekly sales, and holiday data. The evaluation of model performance was conducted through the normalized root-mean-square error (NRMSE).

In another exploration, Kassambara (kassambara, 2018) shed light on the incorporation of interaction effects within a

multiple linear regression framework using the programming language R. He initiated his analysis with a fundamental multiple regression model aiming to predict sales based on advertising budgets allocated to YouTube and Facebook. Kassambara then proceeded to construct an additive model, introducing interaction effects based on two pertinent predictors: the budget designated for YouTube and the budget allocated to Facebook.

III. Methodology

A. Data Set Description

To explore retail performance through Exploratory Data Analysis, a dataset of retail industry is downloaded from Kaggle. (https://www.kaggle.com/datasets/braniac2000/retail-dataset).

The dataset features following several key variables:

- *Order & Customer Information*: Including 'Order ID', 'Order Date', 'Customer Name', and 'Feedback', these fields allow for transaction tracking, temporal sales analysis, and understanding customer interactions.
- *Geographic Data*: 'Country', 'State', 'City', and 'Region' offer insights into sales distribution and regional market trends.
- *Sales Categories*: 'Segment', 'Ship Mode', 'Category', and 'Sub-Category' provide a comprehensive view of market segmentation, shipping logistics, and product classification.
- *Product & Sales Metrics*: 'Product Name', 'Discount', 'Sales', 'Profit', and 'Quantity' are critical for evaluating product performance, pricing strategies, and profitability.

The dataset's comprehensive nature, covering various dimensions of retail operations, sets the stage for an in-depth Exploratory Data Analysis (EDA) to uncover trends and patterns in consumer behavior, sales performance, and operational efficiency.

B. Data Preprocessing & Cleaning

In this phase, we refined our Kaggle-sourced dataset for clarity and analysis suitability. Initially, the dataset was loaded into a Pandas DataFrame named original_data.

- *Column Removal and Renaming*: We began by creating data_1, excluding 'Order ID' and 'Feedback' columns to maintain focus on relevant data. This was achieved using Pandas' drop method. To enhance data accessibility, we standardized column names in data_1, converting spaces to underscores (e.g., 'Order Date' to 'Order_Date'). This resulted in a new DataFrame, data_2, with uniformly formatted column names.
- *Data Type Conversion*: A crucial step was ensuring appropriate data types. Specifically, 'Order_Date' was converted from string to datetime format, enabling more effective time-series analysis. This conversion, executed via Pandas' to_datetime, was verified along with other columns' data types using data_2.dtypes, confirming successful preprocessing.

These steps produced “data_2”, a streamlined dataset poised for in-depth analysis, aligning with our project objectives.

C. Exploratory Data Analysis Techniques

Univariate Analysis: Univariate analysis was performed to understand the distribution and behavior of individual variables within the dataset. We utilized box plots to examine the distributions of 'Sales', 'Profit', and 'Quantity'. These plots revealed significant insights about the skewness and the presence of outliers in these variables. Additionally, a violin plot was created for the 'Discount' variable, revealing its distribution and density across different values.

Bivariate Analysis: In the bivariate analysis, we explored relationships between two variables. A bar chart provided a clear visualization of the variations in sales and profit across different countries, highlighting key differences in market performance. A pie chart was utilized to understand the segment-wise contributions to total sales and profit, offering a perspective on which segments are most influential in the business.

Multivariate Analysis: For multivariate analysis, an area chart was used to observe the year-wise cumulative growth in sales and profit. This visualization allowed us to understand the broader trends over time, revealing the relationship between revenue growth and profitability. We also employed a radar chart to compare the performance of different segments across multiple metrics such as sales, profit, quantity, and discount. This helped in identifying which segments excelled in specific areas. Additionally, a bubble chart was created to depict country-wise sales, profit, and quantity, enabling a comprehensive view of market performance across three dimensions.

D. Statistical Testing Methods

Regression Analysis: We conducted a regression analysis to examine the relationship between 'Discount' and 'Sales'. This analysis helped us understand how discount levels impact sales figures.

ANOVA Test: An ANOVA test was performed to determine if there were significant differences in sales across different product categories, segments, and regions.

Chi-Square Independence Test: The Chi-Square Independence Test was utilized to assess the association between the choice of ship mode and the region.

E. Tools and Libraries Used

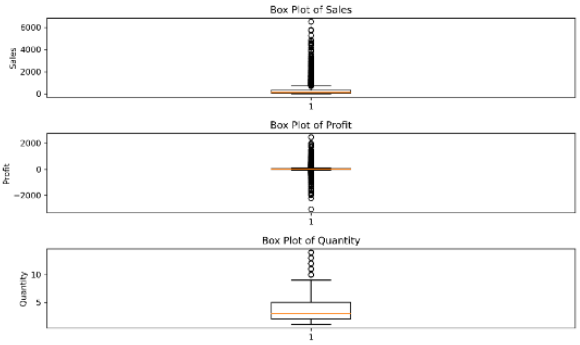
Throughout our analysis, we leveraged various tools and libraries to process and visualize the data. We used Python as our programming language due to its versatility and robust support for data analysis and visualization. Key libraries used included Pandas for data manipulation, Matplotlib and Seaborn for data visualization, and Statsmodels for conducting statistical tests like regression analysis, ANOVA, and Chi-Square tests. These tools were instrumental in allowing us to

efficiently process the data, perform comprehensive analyses, and generate insightful visualizations.

IV. Results

A. Univariate Analysis

1. Box Plot of Sales, Profit & Quantity



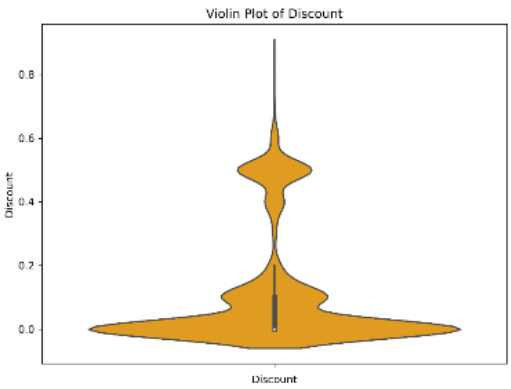
Box Plot Analysis:

Sales:
Display a right-skewed distribution, indicating a majority of transactions are of lower value, with a cluster of much higher-value sales less frequently occurring. A considerable number of outliers suggest sales occasionally include unusually large transactions.

Profit:
The distribution is fairly balanced around the median, but outliers on both the loss and profit sides indicate some transactions are highly profitable or significantly loss-making. The median indicates a general profitability, but the spread of losses and gains needs to be understood for financial health.

Quantity:
Shows a modest skew towards higher values but remains relatively consistent in transaction volume, with outliers indicating infrequent bulk purchases. The bulk of the data points are concentrated in a narrow range, reflecting a stable quantity sold across transactions.

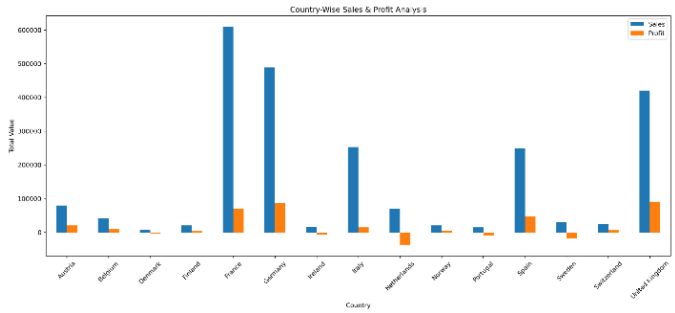
2. Violin Plot (Discount)



Violin Plot Analysis:
The violin plot for "Discount" shows a bimodal distribution with peaks near 0 and 0.2, indicating frequent discounts at 0% and 20%. The distribution is symmetric with a median closer to 0, suggesting that no or lower discounts are more common. The data ranges up to just over 80%, with no significant outliers, implying a consistent discount strategy across sales transactions. The plot's width indicates a higher density of discounts at these common rates, tapering off for larger discounts.

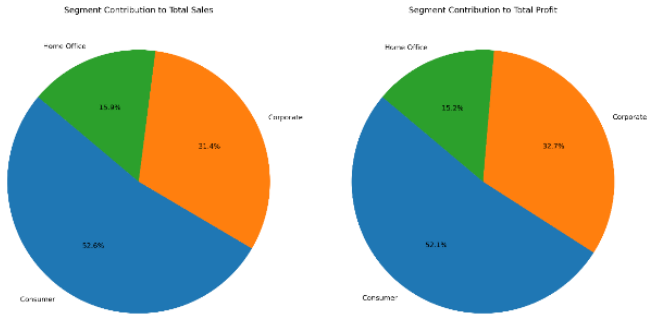
B. Bivariate Analysis:

3. Bar Chart (Country-Wise Sales & Profit Analysis)



Bar Chart Analysis:
The bar chart shows varied sales and profit levels across countries. Germany and France exhibit high sales but lower profits, suggesting high costs or discounts. The UK displays a strong balance with high sales and profits, indicating effective market strategies. Smaller economies like Austria and Belgium have lower sales and profits, possibly reflecting limited market activity. Notably, Germany's high sales don't equate to high profits, hinting at possible market inefficiencies. Spain and Italy show moderate sales with consistent profit margins, suggesting successful pricing strategies. The UK's performance points to a robust market with growth potential.

4. Pie Chart (Segment-Wise Contribution to Sales & Profit)

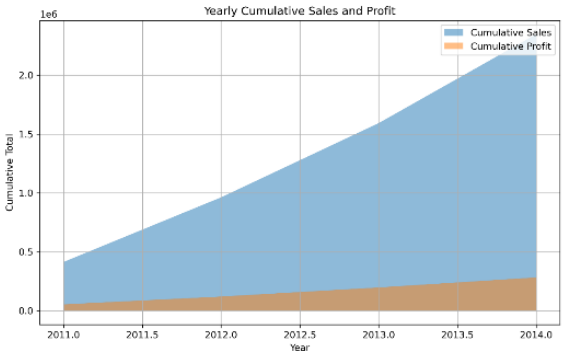


Pie Chart Analysis:
The pie charts show the consumer segment as the leading contributor to both sales (52.6%) and profit (52.1%), highlighting its central role in the business. The corporate segment accounts for a substantial share as well, with roughly a third of sales and profit (31.4% and 32.7%, respectively). The home office segment has the smallest impact, contributing 15.9% to sales and 15.2% to profit. The proportional similarity

between sales and profit across all segments suggests uniform profit margins. This data underscores the consumer segment as the primary target for revenue growth strategies.

C. Multivariate analysis

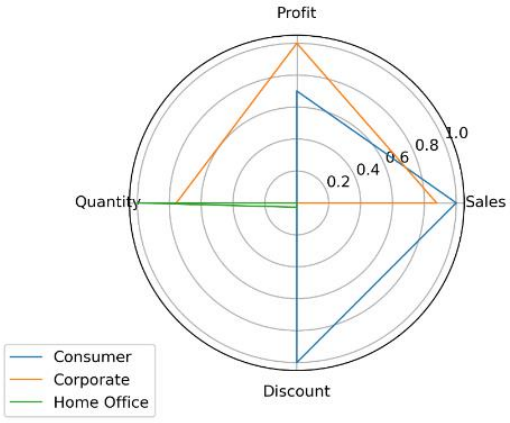
5. Area Chart (Year Wise Analysis of Sales & Profit)



Area Chart Analysis:

The area chart indicates a steady increase in cumulative sales over the years, with a consistent growth in cumulative profit, albeit at a slower rate. This suggests that while revenue is rising, the profit is not increasing at the same pace, potentially due to rising costs or discounts. The gap between sales and profit widens over time, which could indicate a need to improve cost efficiency or adjust pricing strategies to enhance profitability.

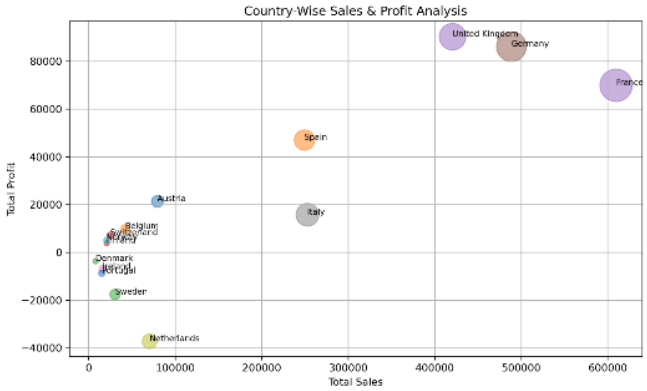
6. Radar Chart (Segment-Wise Contribution to Sales, Profit, Quantity, and Discount)



Radar Chart Analysis:

The radar chart shows segment-wise performance across four metrics. The 'Consumer' segment scores highest on Sales, indicating strong revenue generation. 'Corporate' leads in Profit, suggesting higher efficiency or margin. 'Home Office' excels in Quantity, which may imply bulk purchasing patterns or frequent smaller transactions. However, 'Home Office' also has the highest Discount levels, possibly as a strategy to drive volume. The 'Consumer' segment shows moderate levels of Discounts and Quantity, balancing revenue and sales incentives. 'Corporate' segment is least prominent in Discounts, hinting at less frequent use of pricing strategies to attract sales.

7. Bubble Chart (Country-Wise Sales Profit & Quantity Analysis)



Bubble Chart Analysis:

In the bubble chart, sales and profit data across various countries are represented, with bubble size denoting the quantity sold. The United Kingdom and Germany stand out with high sales and profits, coupled with a large quantity of products sold, as shown by their larger bubble sizes. France follows with strong sales but lower profits, indicated by a smaller bubble. Spain and Italy feature moderate sales and profits with smaller quantities, while the Netherlands is unique, displaying some sales but negative profit, placing it below the zero-profit line. Countries such as Austria, Belgium, and Sweden show lower sales and profit levels with very small quantities sold, represented by the smallest bubbles clustered near the origin.

D. Statistical Testing

1. Regression Analysis: Impact of Discount on Sales

OLS Regression Results

Dep. Variable:	Sales	R-squared:	0.001
Model:	OLS	Adj. R-squared:	0.001
Method:	Least Squares	F-statistic:	5.590
Date:	Sun, 03 Dec 2023	Prob (F-statistic):	0.0181
Time:	16:42:59	Log-Likelihood:	-61182.
No. Observations:	8047	AIC:	1.224e+05
Df Residuals:	8045	BIC:	1.224e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	299.5860	6.321	47.393	0.000	287.195	311.977
Discount	-70.3367	29.750	-2.364	0.018	-128.655	-12.019

Omnibus:	7255.752	Durbin-Watson:	1.951
Prob(Omnibus):	0.000	Jarque-Bera (JB):	288079.634
Skew:	4.304	Prob(JB):	0.00
Kurtosis:	31.020	Cond. No.	5.57

Regression Analysis – Explanation of the result:

The regression output suggests a negative relationship between discounts and sales, with each percentage point increase in discount leading to approximately a 70.34 decrease in sales, which is statistically significant ($p < 0.05$). The model's R-squared is very low (0.001), indicating that the discount

variable alone explains less than 0.1% of the variance in sales, suggesting that other factors are also influential. The F-statistic is significant, but given the low R-squared, the model's explanatory power is weak. The model's diagnostics indicate potential issues with the data's normality and homoscedasticity, as evidenced by the significant Omnibus and Jarque-Bera tests, and the high kurtosis value.

The regression analysis indicates that discounts significantly reduce sales revenue, suggesting that the current discounting strategy may be detrimental to the company's earnings. Given that discounts explain less than 0.1% of sales variance, management should look beyond discounting to other factors that influence sales. They might consider refining the discounting approach, possibly using targeted promotions, and investigate additional variables like customer behavior, market trends, or seasonality. The data also implies potential issues with data distribution, prompting a deeper, more nuanced analysis to develop a robust pricing strategy that balances competitiveness with profitability.

2. ANOVA Test: Differences Across Categories

```
ANOVA for Sales across different Categories:
                    sum_sq      df           F         PR(>F)
C(Category)  2.880180e+08     2.0  721.184167  8.175232e-289
Residual     1.606259e+09  8044.0         NaN            NaN

ANOVA for Sales across different Segments:
                   sum_sq      df         F    PR(>F)
C(Segment)  4.608256e+05     2.0  0.978681  0.375851
Residual    1.893816e+09  8044.0       NaN       NaN

ANOVA for Sales across different Regions:
                  sum_sq      df         F    PR(>F)
C(Region)  2.789983e+05     2.0  0.592467  0.552985
Residual   1.893998e+09  8044.0       NaN       NaN
```

ANOVA Test – Explanation of the result:
The ANOVA test results show that there are significant differences in sales across different product categories, as indicated by a very small p-value ($p < 0.05$), suggesting that category is a strong factor in sales performance. In contrast, the segments and regions do not show a significant difference in sales, with p-values (0.375851 for segments and 0.552985 for regions) well above the 0.05 threshold. The F-statistic for categories is extremely high, confirming a strong effect, while the F-statistics for segments and regions are less than 1, indicating a weak effect on sales. This suggests that product categories are a major determinant of sales variation, whereas segments and regions might not be as influential, or there might be other interacting factors not accounted for in the model.

Given the ANOVA test results, the retail company's management should focus on optimizing their product categories, as they significantly affect sales. Since segments and regions do not show a substantial impact, it may not be cost-effective to tailor sales strategies based on these variables alone. Instead, management should concentrate resources on category-specific marketing and product development, and perhaps explore interactions between categories and other variables like customer demographics or seasonal trends to uncover deeper insights for strategic decision-making.

3. Chi-Square Test: Ship Mode & Region Association

```
Chi-Square Statistic: 7.648283280952471
P-value: 0.26501965528914484
Degrees of Freedom: 6

Expected Frequencies:
[[2674.19063005 1094.20678514 1093.60258481]
 [ 874.52963837  357.83397539  357.63638623]
 [ 234.30794085   95.87249907   95.81956008]
 [ 642.97179073  263.0867404   262.94146887]]
```

Chi-Square Test – Explanation of the result:
The Chi-Square test results show a statistic of 7.6482 and a p-value of 0.2650, which is greater than the conventional alpha level of 0.05. This indicates that there is no significant association between the choice of ship mode and the region; in other words, ship mode is likely independent of the region where the goods are being shipped. The expected frequencies table supports this, as there doesn't seem to be a large discrepancy between expected and observed frequencies.

For the retail company's management, this could mean that shipping strategies do not need to be tailored for different regions, potentially simplifying logistics and distribution plans. Centralized shipping operations could be considered, as customer preference for shipping methods appears consistent across regions. This uniformity could allow for bulk transport deals and streamlined shipping operations, leading to cost savings and operational efficiencies. Management might also explore other areas for optimization, given that shipping mode does not influence regional sales.

V. Conclusions

The univariate, bivariate, and multivariate analyses uncovered significant patterns within the retail dataset. Sales exhibited a right-skewed distribution, while profit distribution was balanced but showed extremes, suggesting the need for a detailed understanding of transaction profitability. The impact of discounts on sales was examined through regression analysis, revealing a statistically significant negative correlation. Product categories were identified as a major determinant of sales variation, emphasizing their significance in retail strategies. The Chi-Square test indicated the independence of ship mode choice from regions, offering insights into potential operational optimizations.

Retail decision-makers can leverage the study's findings to refine pricing strategies, optimize product categories, and streamline shipping operations. The negative correlation between discounts and sales implies a need for a nuanced approach to discounting, potentially through targeted promotions. The focus on product categories as a key driver of sales variation suggests the importance of tailored marketing and product development strategies. The independence of ship

mode and regions implies potential cost savings through centralized shipping operations.

REFERENCES

[1] Retail dataset. Kaggle. Retrieved from https://www.kaggle.com/datasets/braniac2000/retail-dataset

[2] DeHoratius, N., Musalem, A., & Rooderkerk, R. (2023, February 27). Why retailers fail to adopt advanced data analytics. Harvard Business Review. https://hbr.org/2023/02/why-retailers-fail-to-adopt-advanced-data-analytics

[3] Agrawal, R. (2022, August 31). Exploratory data analysis using data visualization techniques! Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/exploratory-data-analysis-using-data-visualization-techniques.

[4] Jeswani, R. (2021, December). Predicting Walmart sales, exploratory data analysis, and Walmart sales dashboard. B. Thomas Golisano College of Computing and Information Sciences, Rochester Institute of Technology. https://www.rit.edu/ischoolprojects/sites/rit.edu.ischoolprojects/files/document_library/Rashmi_Jeswani_Capstone.pdf

[5] Harsoor, A. S., & Patil, A. (2015). Forecast of sales of walmart store using big data applications. International Journal of Research in Engineering and Technology. https://ijret.org/volumes/2015v04/i06/IJRET20150406008.pdf

[6] Crown, M. (2016). Weekly sales forecasts using non-seasonal arima models. http://mxcrown.com/walmart-sales-forecasting/

[7] kassambara. (2018). Interaction effect in multiple regression: Essentials statistical tools for high-throughput data analysis (sthda). http://www.sthda.com/english/articles/40-regression-analysis/164-interaction-effect-in-multiple-regression-essentials/