Business Analytics Assignment

A Comprehensive 3-way Analysis, based on 2 Models.

Business Analytics Course

Prof. Drª Conceição Silva
Prof. Dr. João Nuno Gonçalves

22/03/2024

Francisco Nieto Guimarães – 356321002

Inês Sousa e Silva – 355423027

José Miguel Guedes - 355423064

Contents

1. Introduction

During the Business Analytics course, lectured on the 2nd semester of the 1st year in the Master in Management, the class was given tools, procedures and both general and in-depth knowledge regarding the appropriate interpretation and management of data analysis, within the business context. The proposed report aims to affirm concretely the previous statement, in order to further solidify the usage of data as an asset which grants competitive advantage via leveraging value from itself.

Presented with various datasets containing a wide variety of information, from numerous business areas, the class was entrusted with the proper application of the given instruments and methodologies in order to apply an accurate and conscious use of data, to achieve optimal decision-making. Adding to this, the core of the report focuses on the 3-way analysis taught during the course – Descriptive, Predictive and Prescriptive -, on which the group shall pivot towards the 'Online Store Customer Data' dataset for the Descriptive and Predictive component, while the Prescriptive segment will refer to the 'Moore Pharmaceuticals Model' problem.

The fundamental aspect of the group's approach is to present both the interpretation of the data and their respective conclusions in an accessible and intuitive way, in order for any reader to easily assess the informational contents and grasp them for immediate usage, while simultaneously opening a path for multiple applications of the contents presented.

With the goal of providing a holistic vision of the proposed datasets, for the Descriptive and Predictive analyses two models were created, with the first eliminating observations based on an empty-cell basis, meaning that rows with blank values on certain categories were deleted, while on the second model the blank values were filled in, utilizing median, mean values or other methods, and maintaining the same initial number of observations. On each of the chapters presented below, an in-depth explanation shall be provided.

The presented report commits itself in providing a better perception of customer behavior, pattern identification and a clearer image that culminates in a superior performance of the entities under analysis.

2. Descriptive Analysis

The descriptive analysis was made using the "Kaggle online_store_customer_data.xlsx" original file, which compiles information regarding a set of variables regarding a certain e-commerce platform, focused on the United States' market throughout the 2019-2021 financial years.

The dataset that was used to carry out this assignment presented several problems, ranging from missing values and records that were poorly formatted, such as the date of the transaction, to completely non-existent values filling the cells.

In this analysis, 2 models were created, as previously mentioned in the introduction: the first model we will focus on eliminated every single row containing a blank cell, while the second model substituted the missing values using the mode, mean or median, depending on how efficiently and logically they would fit the specified category. For a clearer academic experience, these models will be referred to as 'Restrictive Model' and 'Logical Model'. No substitutions were made on the Restrictive and the remaining number of observations totalled 2044, while on the Logical one the number of observations came up to 2270.

The core reasoning behind this decision came from the group's intention to see if the 10% cut would somehow influence, positively or negatively, the overall data distribution, the inherent results derived from the alterations of the base information and the possible shocks it could have on future observations, or if this distinction would not have any significant impact.

2.1 Data Preparation and Database Organization

The first step for the process of making any Descriptive analysis lies in confirming if the data available is ready to be worked on or if it lacks determined requisites and, specifically, if it's well distributed and organized, as well as making sure that all errors and missing data, noisy information and inconsistent data are sorted in order to avoid distortions in the study, granting precision and interpretation accessibility.

On the sheet named "Base Data", we can simultaneously access the informational bulk of each observation and the relevant categories from which the relevant pivot tables shall be built from. They contain all the imperative data distribution throughout various categories, as well as the implementation of this approach enables a more thorough examination of both trends and patterns within the past data.

2.2 Logical Model

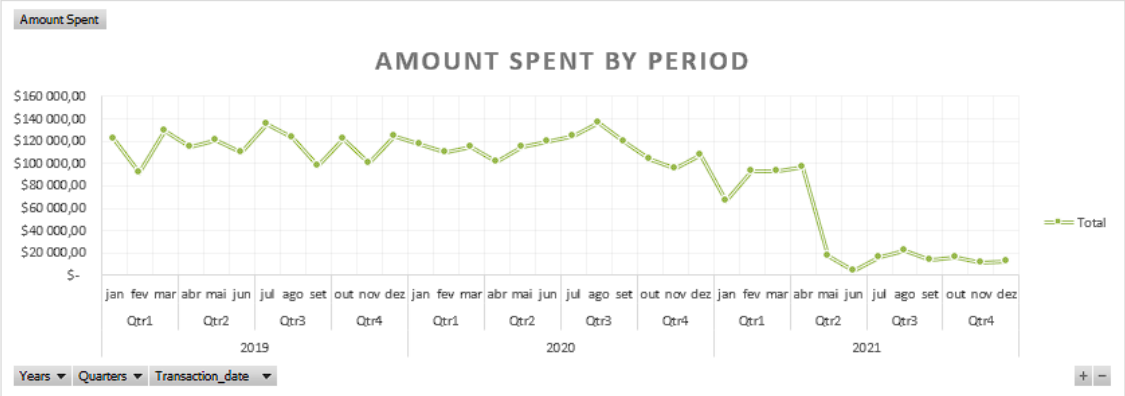
The attribute "Employees_status" was filled through the mode, the "Age" by using the median and the "Referral" by utilizing the forward fill method. Through the removal of the records that had non-values on the "Amount_spent" column, in order not to bias the predictive analysis, since this attribute was used to make a forecast. At last, the creation of a new value to fill the "Gender" attribute as "Unknown" was inserted, and as a consequence it transformed all the values on "Transaction_date" column to a date format. The process as a whole was almost completely prepped via the assistance of the programming language Python.

2.2.1 Time Period

In the "Period" sheet, the analysis conducted by both the pivot table and the graph was focused on the relationship between the amount spent and the specific time periods - years, quarters and months.

It is possible to deduce that during 2019 and 2020, the amount spent in the store ranged between $100,000 and $140,000, with the peak spending having occurred in July 2019 and August 2020, both approximately reaching $140,000. However, in 2021, there was a notable sharp decrease in expenditure, particularly noticeable from May through December, with total spending being decreased significantly to as low as $20,000.

A pivot table showing the relationship between the amount spent and the number of transactions was created, and it displays an increase of store prices going up from 2019 to 2021, as a consequence, as the ratio of amount spent to the number of transactions raised even further during this period.

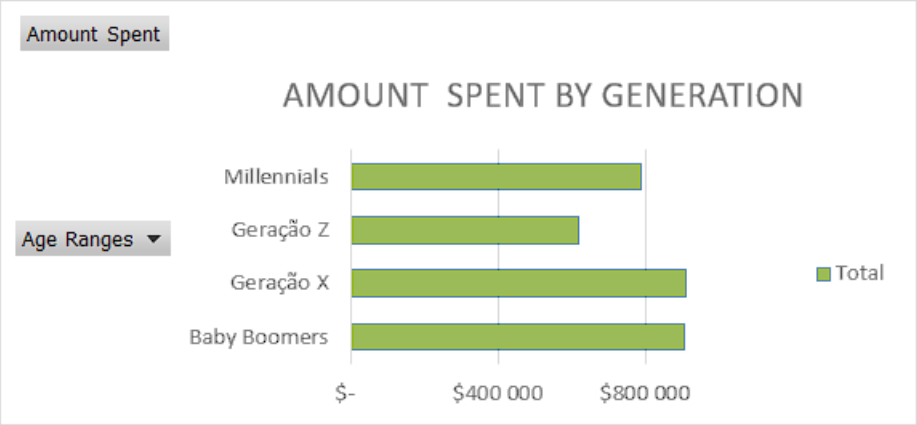


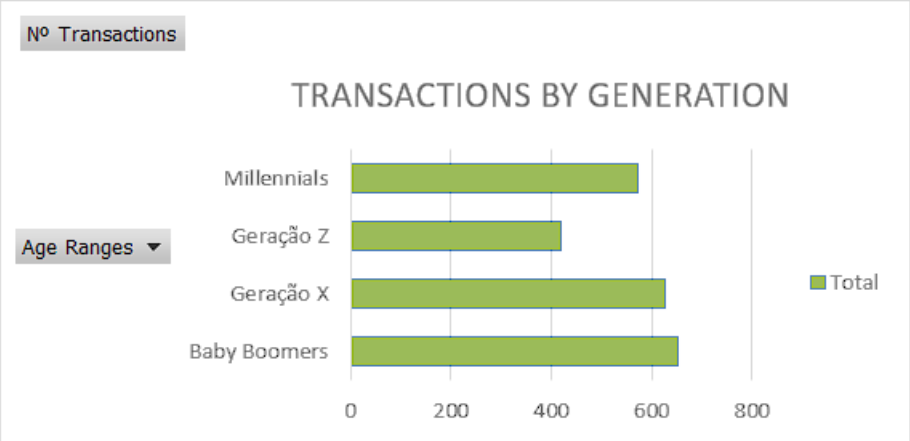
Row Labels	Count of Transactions	Sum of Amount Spent	Amount Spent / Transactions
2019	990	1 390 702	1 405
2020	958	1 365 318	1 425
2021	322	463 799	1 440
Grand Total	2 270	3 219 819	1 418

2.2.2 Generation

In this part of the analysis, the ages provided in the datasheet were associated and sorted with generations, in order to enhance our analysis. From this, the conclusion that Baby Boomers (>59 years) and Generation X (44 to 59 years) are simultaneously the top buyers and spenders of the platform, while Generation Z (28 to 43 years) stand out for the opposite.

Comparing the two graphs, the amount spent by Baby Boomers and Generation X is quite similar, but Baby Boomers have a higher number of transactions. Therefore, it's easy to infer that Generation X's purchases hold higher value compared to the Baby Boomers'.

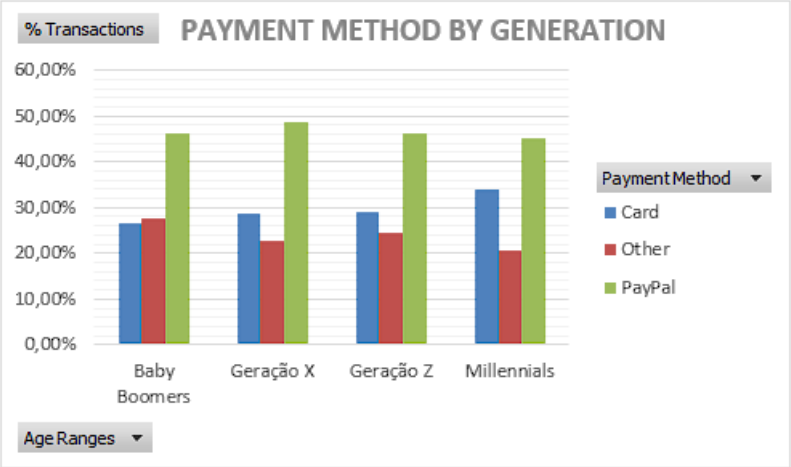




2.2.3 Payment Method by Generation

In the analysis focusing on generations and their payment preferences, some valuable insights were uncovered. Across all generations, PayPal emerged as the most popular payment method, constituting between 45% to 50% of transactions. The second most preferred method is (Credit/Debit) Card, utilized by three generations: Generation X (44 to 59 years), Generation Z (28 to 43 years), and Millennials (<27 years old).

However, for Baby Boomers, the second most used payment method falls under the "Others" category, likely indicating cash payments. This highlights a notable difference in payment behavior among different generations.

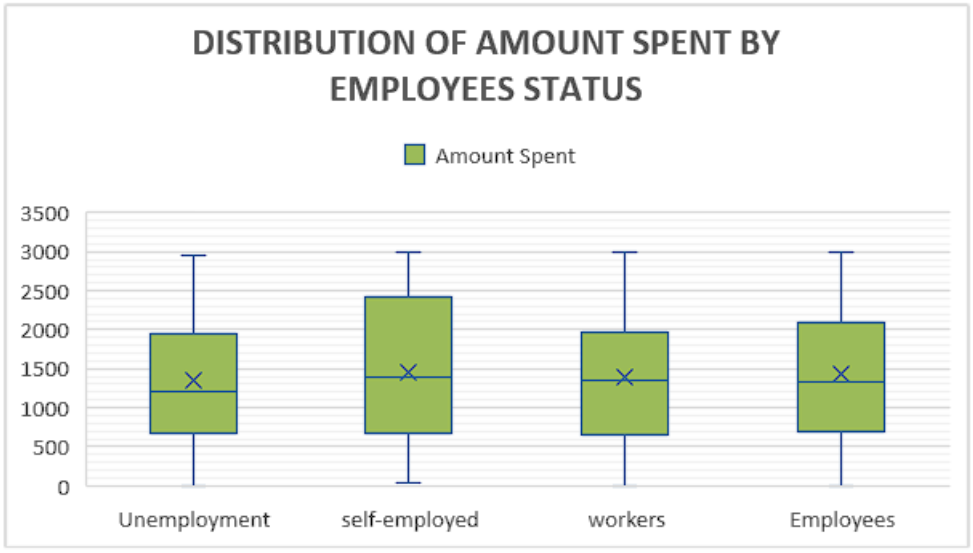


2.2.4 Employee Status

In this box plot, it is observable that individuals categorized as unemployed, workers, and employees exhibit a fairly similar distribution of amount spent. In contrast, self-employed individuals stand out due to their distinct distribution pattern. None of the characterization's present outliers, having the same maximums and minimums.

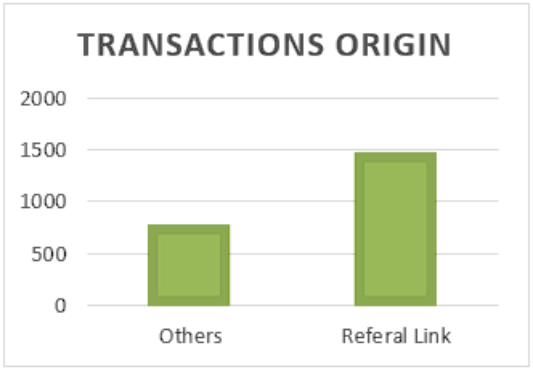
Approximately 50% of self-employed individuals who make purchases at the store spend approximately between 700 to 2400 in their transactions. On the other hand, 50% of individuals with other employment statuses spend between 700 to only 2000. This indicates a notable difference in spending behavior based on employment status.

In relation to the median, employee's status self-employed, workers and employees have a median identical to the average, with only unemployment having a lower median. This reflects a positive skewness on the part of unemployment, with the mean being higher than the median and in turn the median being higher than the mode. The remaining employment statuses have medians equal to or close to the mean, that is, there is a symmetric distribution, so the mean, mode and median have equal values.



2.2.5 Referral Link

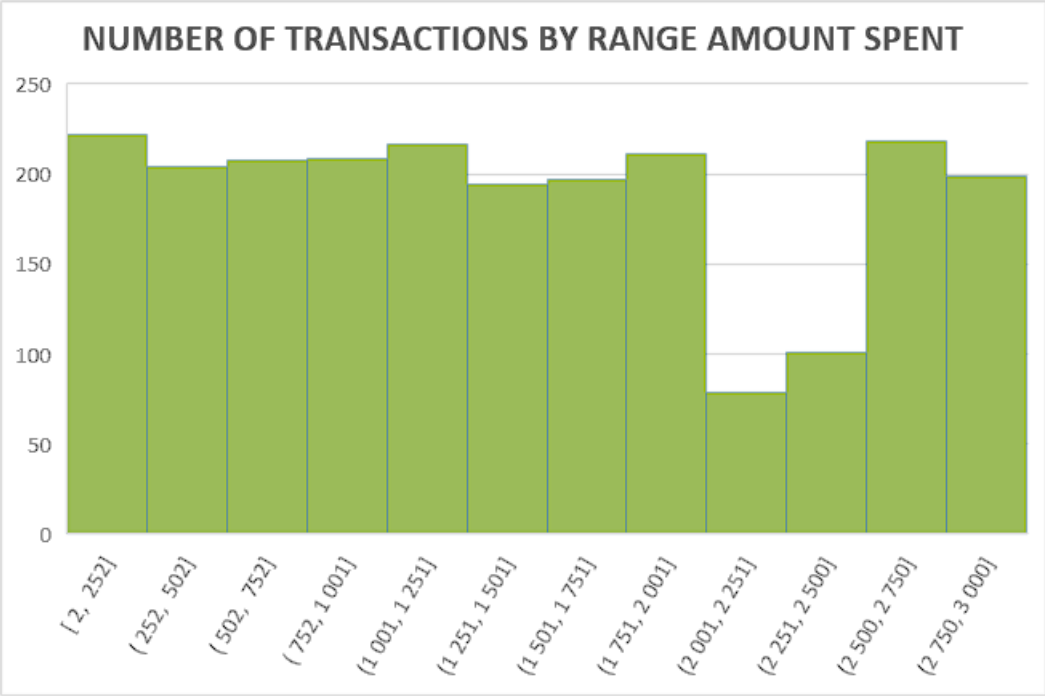
In the "Referral Link" worksheet, the presence of a simple pivot table and a column chart provide valuable information to the store. Out of the total 2270 transactions, approximately 1481 transactions were made through customers clicking on the referral link provided by the store. This means that only 35% of transactions are conducted through other methods.



2.2.6 Range Amount Spent

In the figure below, a histogram displays the distribution of amount spent in transactions across 12 value ranges.

One notable aspect of this analysis is the prevalence of purchases in both the lower value range (between $2 to $252) and one of the higher value ranges ($2500 to $2750). It's also intriguing to note that there are only about 180 transactions between $2000 and $2500, which is the lowest count compared to all other value ranges. There are more transactions either below $2000 or above $2500.



2.2.7 Other Indicators

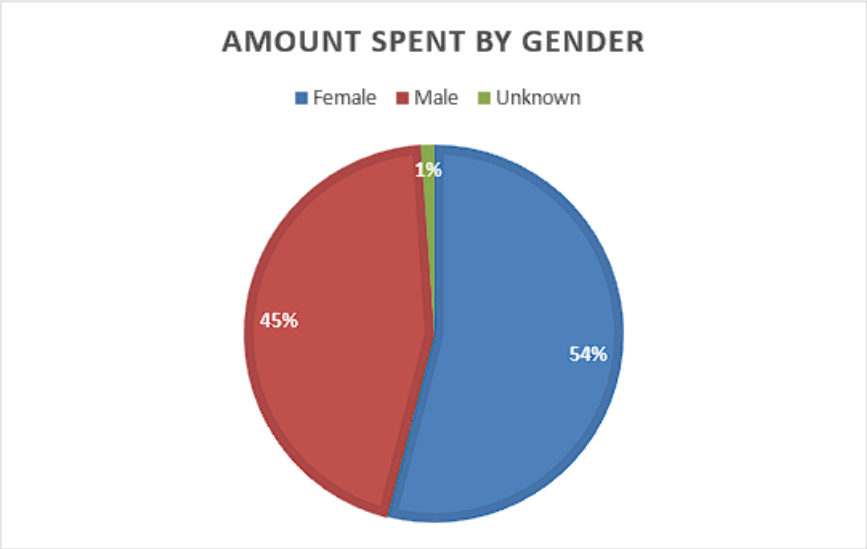
Other pivot tables and charts created to draw conclusions based on different analysis categories shall be referenced throughout this current chapter.

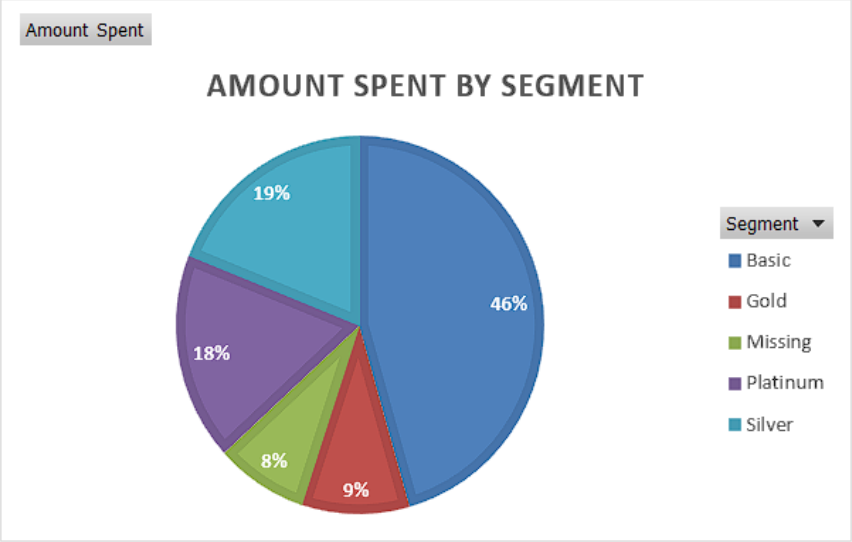
In the first pie chart below, a comparison of the spending in our store by gender was generated, displaying a 54% of the spending coming from women, a percentage that, based on the analysis, doesn't warrant focusing all efforts solely on this audience.

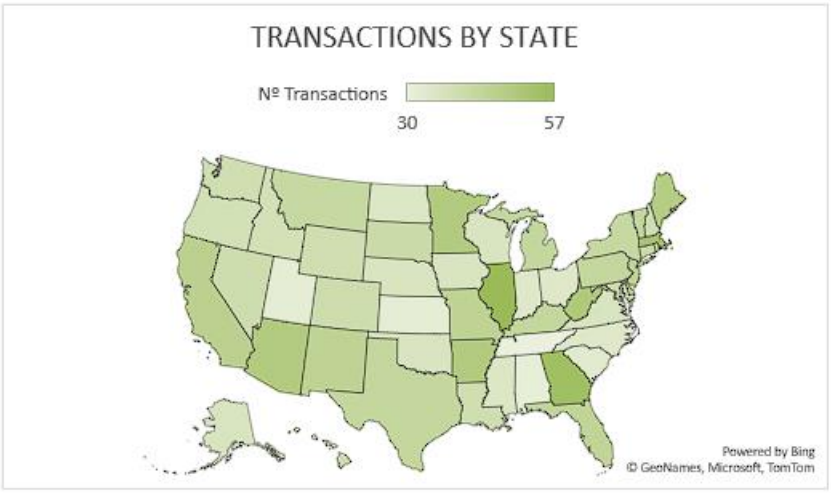
In the second pie chart, a study on the spending by product category reveals that only 8% and 9% of the amount spent corresponds to the "Missing" and "Gold" segments, respectively.

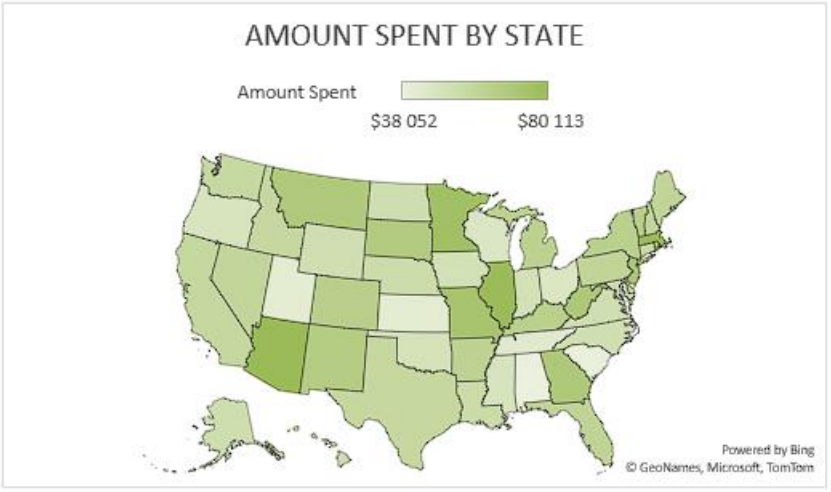
Despite all this, nearly 50% of the spending corresponds to the "Basic" segment, indicating its importance for the store.

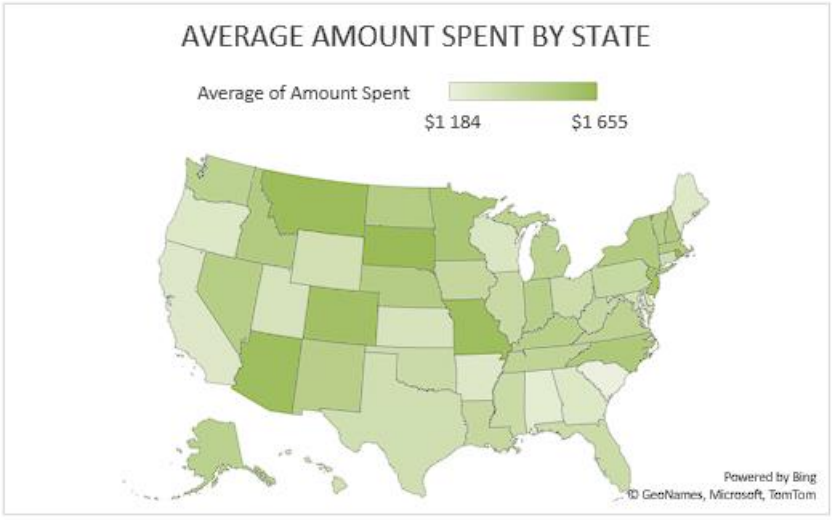
Lastly, three maps illustrating the states where the store operates were produced. One map shows the number of transactions per state, another shows the amount spent per state and lastly, an average spending per state. Both maps yield some similar conclusions, notably that Illinois is the state with the highest spending and transactions in the store. On another hand, Nebraska has fewer transactions (36) but ranks higher in amount spent compared to nine other states, indicating that Nebraska customers have higher transaction values.






TRANSACTIONS BY STATE
Nº Transactions
30
57
Powered by Bing
© GeoNames, Microsoft, TomTom


AMOUNT SPENT BY STATE
Amount Spent
$38 052
$80 113
Powered by Bing
© GeoNames, Microsoft, TomTom


AVERAGE AMOUNT SPENT BY STATE
Average of Amount Spent
$1 184
$1 655
Powered by Bing
© GeoNames, Microsoft, TomTom

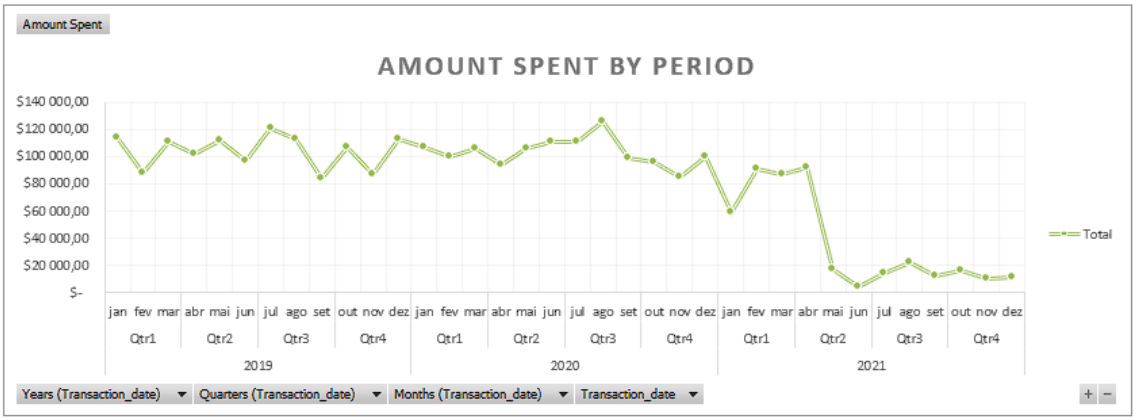
2.2.8 Dashboard

A dashboard was created to dynamically and interdependently display all the charts and analyses mentioned above. The dashboard features slicers for Segments, Gender, Age Ranges (Generations), States, and also includes a timeline spanning the three years' months. This setup allows for data filtering based on the specific criteria required for a more focused analysis.

2.3 Restrictive Model

2.3.1 Time Period

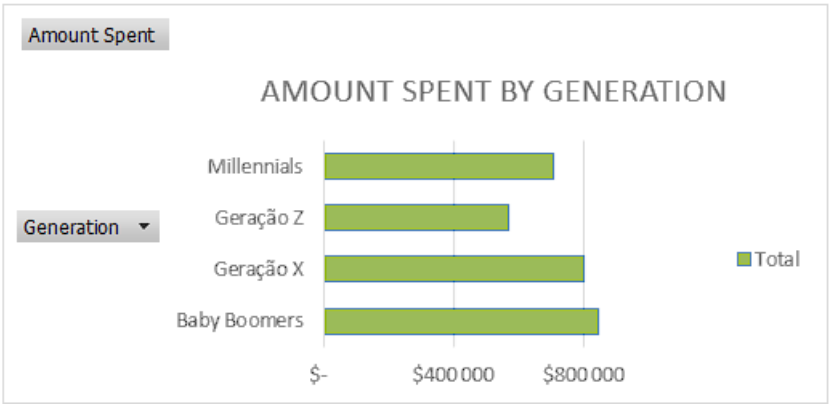
Similar conclusions from the Logical Model can be drawn, most easily identifiable the spending that occurred during 2021 and the overall trendline of general amount spent. Despite all this, some distinctions can be immediately inferred, for example, the fact that 2019 had a drop of nearly 143k$ and 2020 a decrease of 127k$, which overall, represents a 10% reduction from the first model.

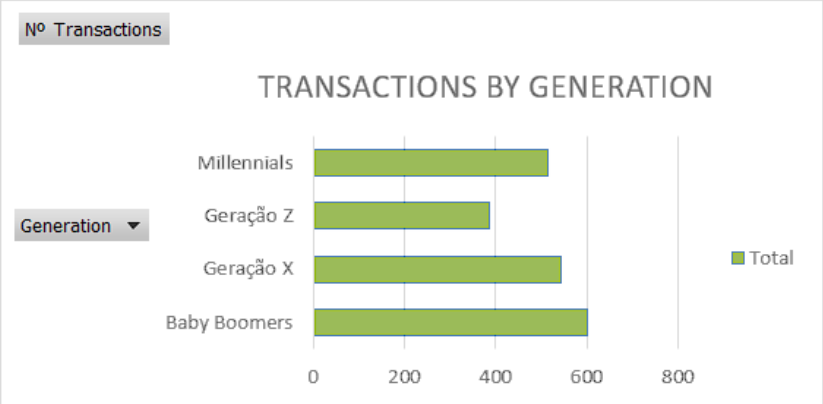


Period	Amount Spent	Count of Transactions	Amount Spent / Transactions
2019	$ 1 247 649,75	879	1419
2020	$ 1 237 644,29	864	1432
2021	$ 436 525,19	301	1450
Grand Total	$ 2 921 819,23	2044	1429

2.3.2 Generation

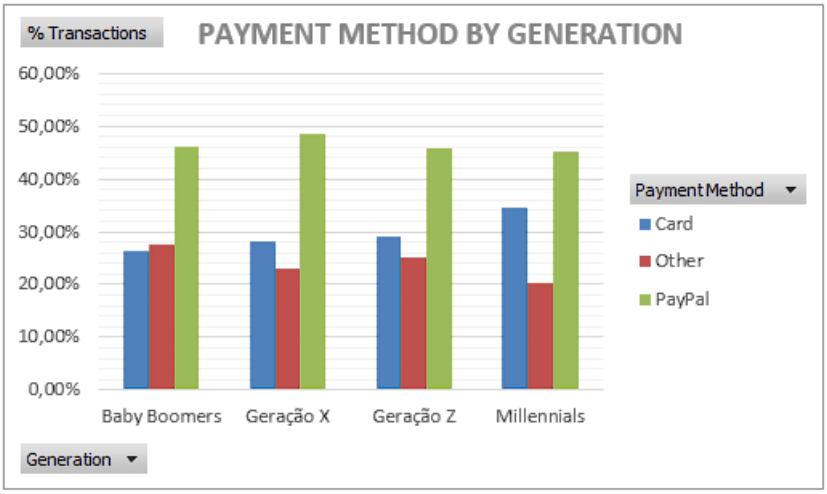
The trend established of Baby Boomers (>59 years) and Generation X (44 to 59 years) being the top buyers and spenders in the store remains consistently, with Generation Z (28 to 43 years) standing out, likewise, for the opposite.





2.3.3 Payment Method by Generation

The exact same consumer behaviors and patterns persist on the payment method topic, with each generation's respective preference prevailing over the remaining ones, with no notable exceptions or distinctions

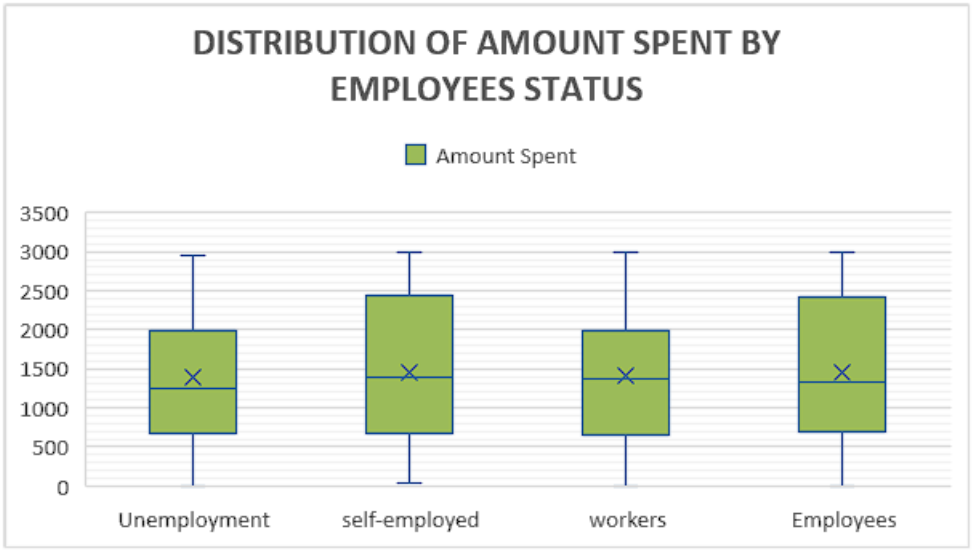


2.3.4 Employee Status

In this box plot, we can observe that individuals categorized as unemployed, workers, and employees exhibit a fairly similar distribution of amount spent. In contrast, self-employed individuals stand out due to their distinct distribution pattern.

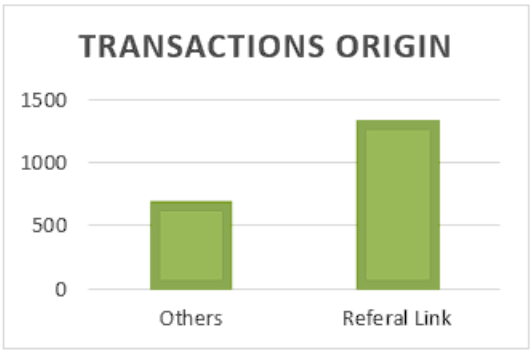
Approximately 50% of self-employed individuals who make purchases at the store spend approximately between 700 to 2400 in their transactions. On the other hand, 50% of individuals

with other employment statuses spend between 700 to only 2000. This indicates a notable difference in spending behavior based on employment status.



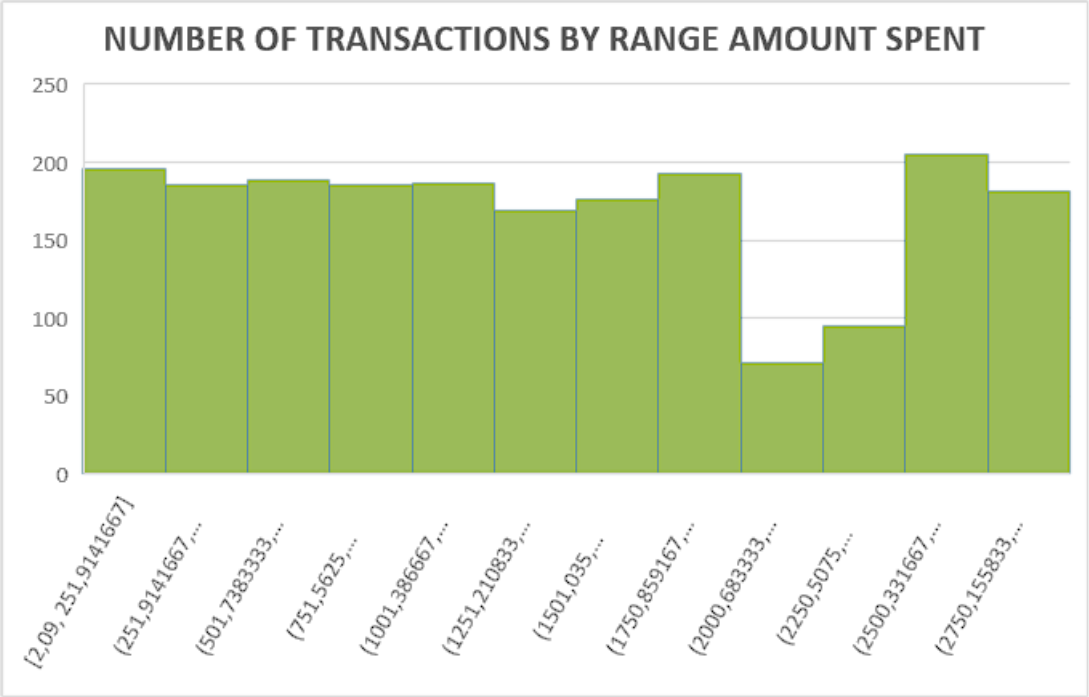
2.3.5 Referral Link

Out of the total 2044 transactions, approximately 1342 transactions were made through customers clicking on the referral link provided by the store, with the same 35% utilizing other paths to reach the store – previously concluded on the Logical Model -, thus establishing that this category holds little to no significant change in the overall analysis.



2.3.6 Range Amount Spent

Unlike the first model, due to the cut in total observations, a relevant decrease in all columns is observable, with no exceptions being made, evident by the fact that fewer observations correlate directly to lower transaction numbers total. The most noticeable drop in range amount spent is found from $1001 upwards to $3000.

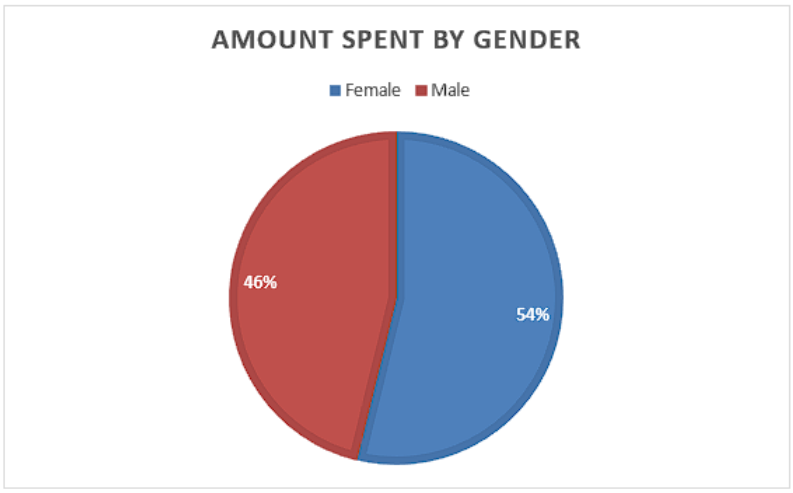


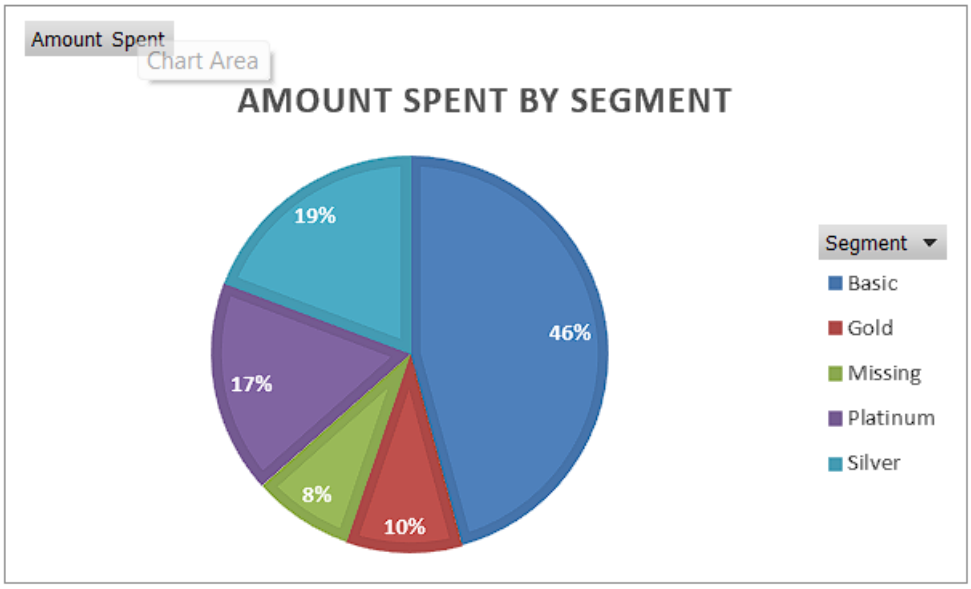
2.3.7 Other Indicators

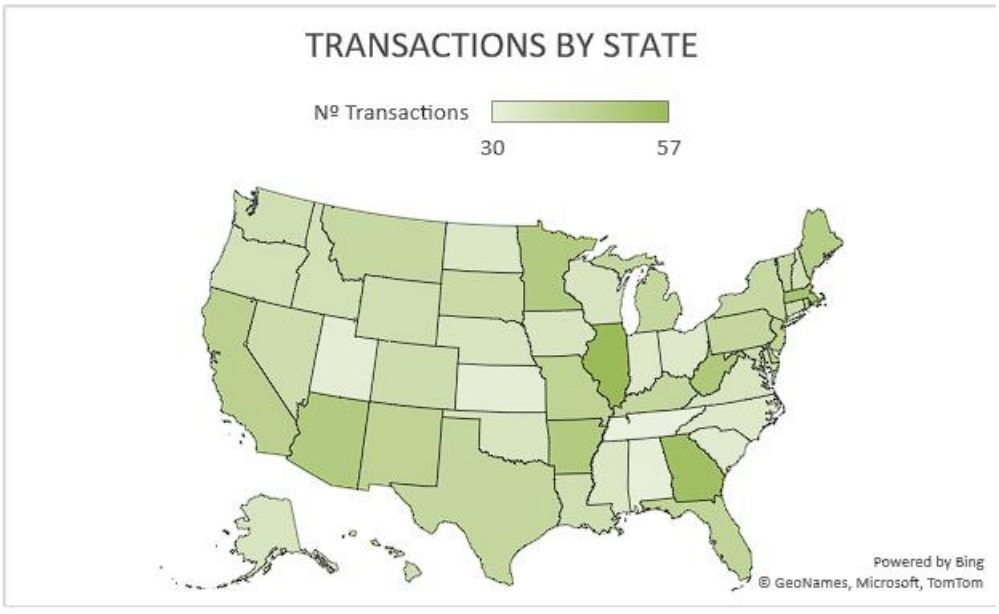
In the first pie chart below, due to the elimination of the Unknown values, the amount spent by Males was increased by 1%, with no additional changes to the Female proportion.

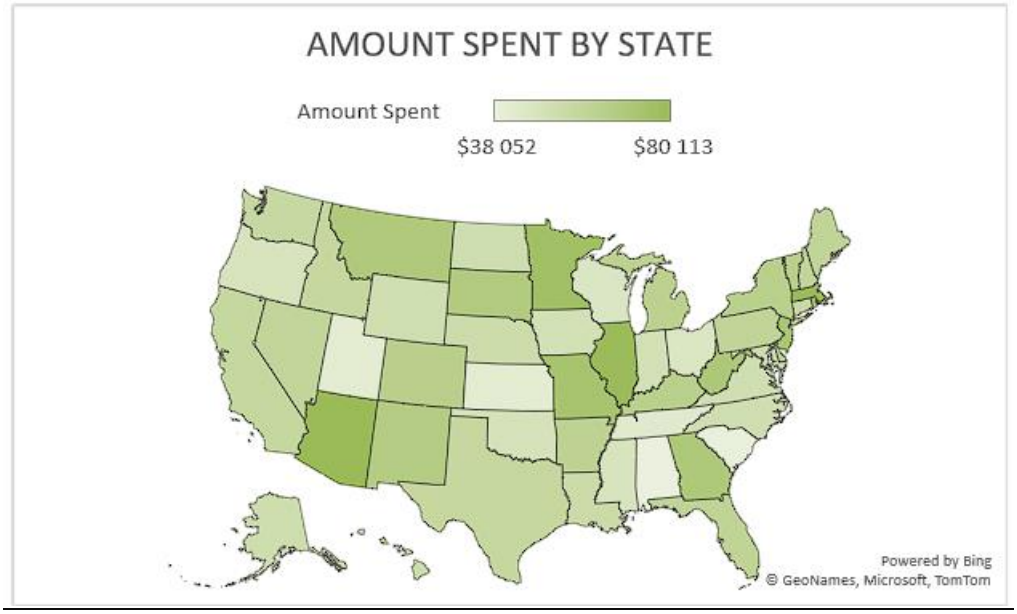
In the second pie chart, the “Gold” segment had its total raised to 10%, while “Platinum” dropped to 17%.

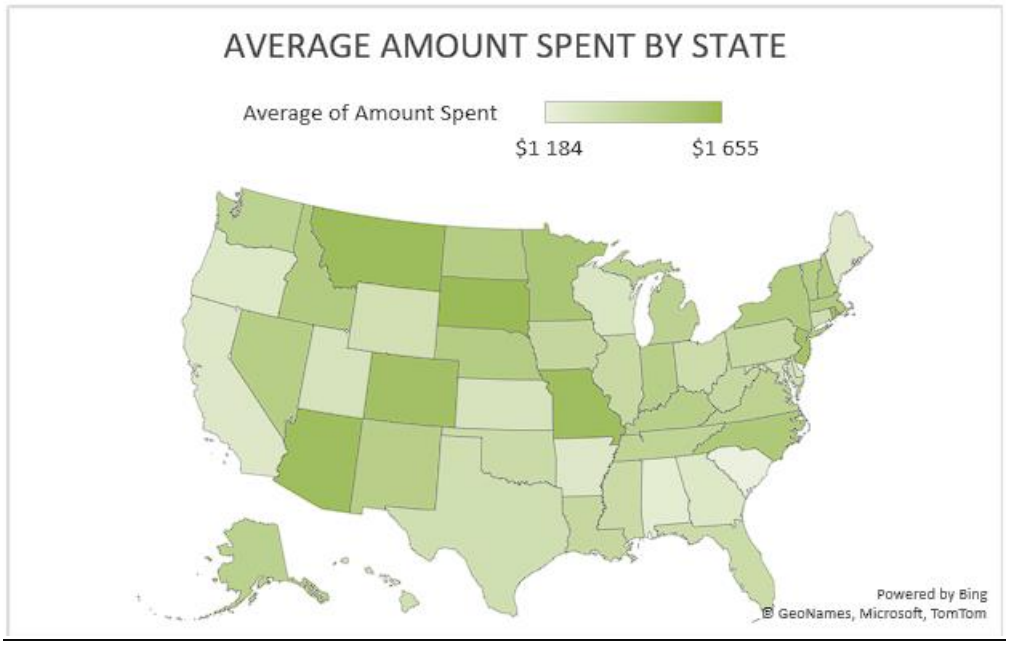
Regarding each state, similar conclusions from the first model were assed, displaying low significance in this stage of the analysis.


AMOUNT SPENT BY GENDER
Female
Male
46%
54%


Amount Spent
Chart Area
AMOUNT SPENT BY SEGMENT
19%
46%
17%
8%
10%
Segment
Basic
Gold
Missing
Platinum
Silver


TRANSACTIONS BY STATE
Nº Transactions
30
57
Powered by Bing
© GeoNames, Microsoft, TomTom


AMOUNT SPENT BY STATE
Amount Spent
$38 052
$80 113
Powered by Bing
© GeoNames, Microsoft, TomTom


AVERAGE AMOUNT SPENT BY STATE
Average of Amount Spent
$1 184
$1 655
Powered by Bing
© GeoNames, Microsoft, TomTom

2.3.8 Dashboard

3. Predictive Analysis

3.1 Predictive Analysis - Logical

Regarding the predictive analysis two methods were utilized with the goal of predicting the amount spent in January, February and March of 2022, via the *linear regression's models* and *Holt's linear trend method*.

3.1.1 Linear Regression Models

Starting with *Multiple Linear Regression*, variables as "Gender", "Marital Status" and "Employee Status" were changed to *dummy variables* in order to be used, together with "Age", as independent variables. The reasoning behind the selection of these variables for the regression lies in the fact that, based on the descriptive analysis, possible patterns are notable

in a given gender, marital status or employment situation in regards to possible higher consumption and spending.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,046715227
R Square	0,002182312
Adjusted R Square	-0,001789714
Standard Error	878,9051233
Observations	2270

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	9	3821574,499	424619,3888	0,706741	0,7032565
Residual	2262	1747336676	772474,2157		
Total	2271	1751158250			

As can be seen from the *ANOVA* table, the *Significance F* presents a value greater than 0.05 and therefore the null hypothesis cannot be rejected, in which this regression, in general, is neither useful nor credible for predicting values of "Amount Spent" and that there is no relationship between these variables. Furthermore, the coefficient of determination has a value extremely close to zero, concluding that the regression cannot predict the values. This result is explained by the high value of the *Residual Sum of Squares* (1 749 456 494) and the high percentage that this value has in relation to the *Total Sum of Squares* (1 751 158 250). This phenomenon is a consequence of the abysmal difference between the observed values of the dependent variable and the predicted values from the regression model.

$$Significance\ F = 1 - \frac{SS\ residual}{SS\ Total}$$

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*
Intercept	1506,701005	68,77936499	21,90629422	1,2575E-96	1371,823757	1641,57825	1371,823757	1641,578254
Age	-1,145308172	1,025242858	-1,117109144	0,2640663	-3,155823038	0,86520669	-3,15582304	0,865206694
Marital Status	-9,218765299	37,47958693	-0,245967633	0,8057296	-82,71673321	64,2792026	-82,7167332	64,27920261
Unemployment	-107,6702395	72,25527567	-1,49013672	0,13632771	-249,3637951	34,023316	-249,363795	34,02331604
Self-Employed	0	0	65535	#NUM!	0	0	0	0
Workers	-64,12169254	52,98663068	-1,210148517	#NUM!	-168,0291793	39,7857942	-168,029179	39,78579419
Employees	-22,87675555	51,04881954	-0,448134859	0,65409879	-122,9841689	77,2306578	-122,984169	77,23065779
Male	0	0	65535	#NUM!	0	0	0	0
Female	20,71705066	37,27337813	0,555813605	#NUM!	-52,37653907	93,8106404	-52,3765391	93,81064038
Unknown	-66,0525858	174,592023	-0,378325336	0,70522443	-408,4298625	276,324691	-408,429862	276,3246908

All variables are not essential in the regression, since they have a *p-value* greater than 0.05, and the null hypothesis is not excluded, and these variables have a value of 0. Furthermore, they have high value in the standard error, that is, the predicted values are quite far from the population. Variables *like Male, self-employed and workers* have an error in p-value and zero value in coefficient and other aspects of regression analysis since Excel's data analysis tool does not consider them significant or contributory in the regression.

This is clearly an error which demonstrates that this method is not suitable for predicting "Amount Spent" since, despite there being a mere indirect relationship between "Gender", "Employee Status" and "Marital Status", there is no cause-effect relationship strong enough to

predict o "Amount Spent". Furthermore, this *Multiple Linear Regression* is quite weak due to the use of variables that were categorical and were transformed into *dummy variables*.

In the case of using a *Simple Linear Regression*, with age being the only independent variable, despite the variables already being related and in general the regression is useful and significant but there would still be no credibility in predicting "Amount Spent" values since the coefficient of determination is very close to zero.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,023277685
R Square	0,000541851
Adjusted R Square	0,000101172
Standard Error	878,4630097
Observations	2270

The standard error shows that the estimates are not very precise, and the estimated coefficient is around 878 units of the real population coefficient.

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	948866,2064	948866,2064	1,229583486	0,267605623
Residual	2268	1750209384	771697,2593		
Total	2269	1751158250			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	1471,469498	51,26904416	28,70093487	8,3028E-155	1370,930363	1572,008632	1370,930363	1572,008632
Age	-1,133940178	1,022612584	-1,108865856	0,267605623	-3,139294202	0,871413845	-3,139294202	0,871413845

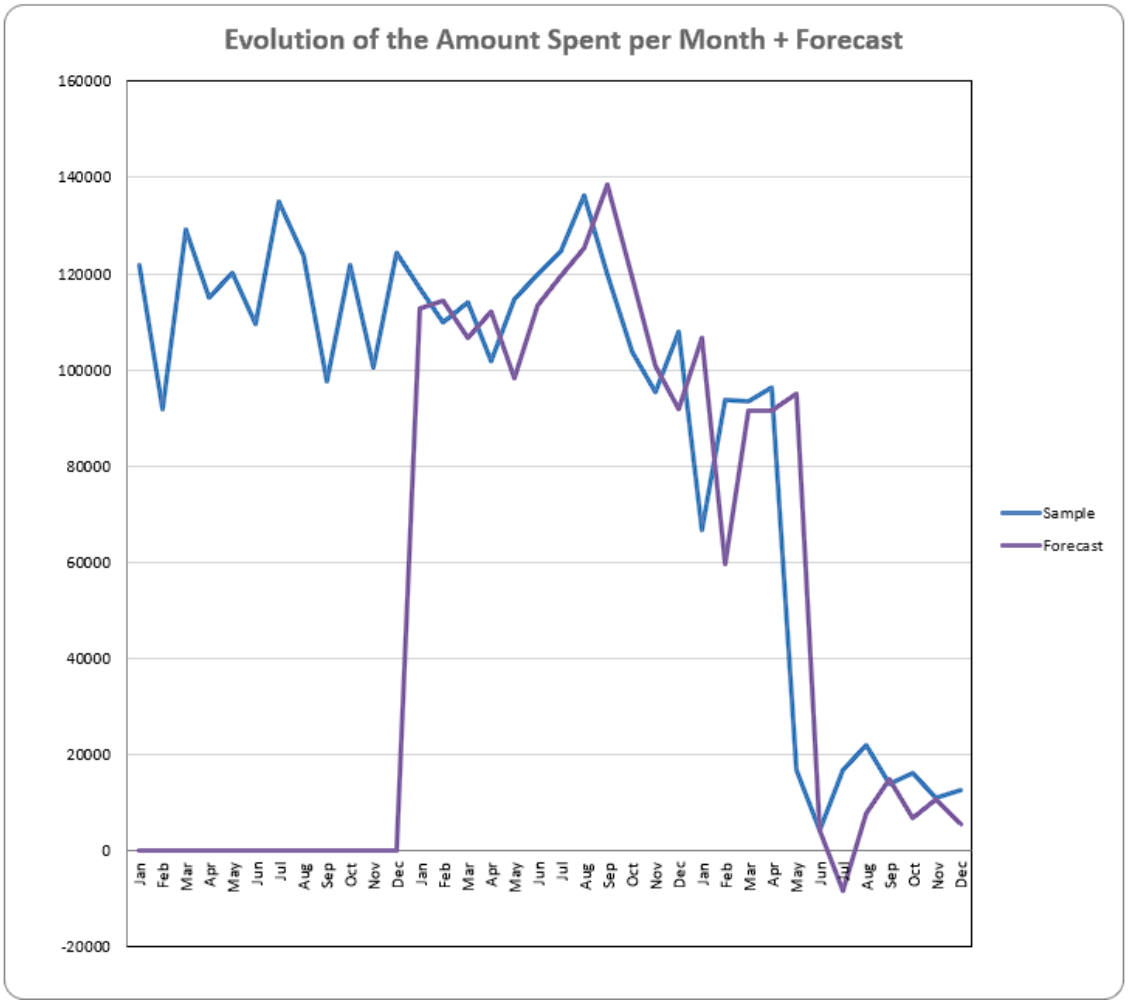
$$Amount\_Spent = 1471.47 - 1.13 * Age$$

In conclusion, a *Linear Regression Model*, whether simple or multiple, is not an ideal method for predicting "Amount Spent" values since there is no strong causal relationship between the independent variables and the dependent variable.

3.1.2 Holt's Linear Trend Method

Firstly, this method was chosen since there is a level and trend over the months used for the analysis.

As is possible to see from the graph, there is a tendency for the amount spent to decrease over time and there is no seasonality, as there is no repetitive cycle of increase or decrease.



The "Transaction Date" was transformed into total "Amount Spent" per month in order to predict the total "Amount Spent" in the months of January, February, and March 2022. Initially, Alpha and Beta values of 0.5 were given to calculate the level (St) and the trend (Tt). To initialize the method calculation, it began in January 2020, in which the initial St was the average of the previous 12 months, and the Tt was -3172.2, referring to the trend line.

After that, an error measure, MAPE, was calculated. This presented a value of 85.80%, concluding that the model is not accurate in predicting values, since the difference, in perspective, between the predicted values and the actual values of the sample are quite high.

Months	Forecast
Jan	11887.3596
Fev	13066.5486
Mar	14245.7376

MAPE	85.80%

Using the solver tool, with the objective of minimizing the MAPE by changing the alpha and beta, the best optimization achieved was with an alpha of one, that is, it gives more focus to more current events, and a beta of 0.145, so to minimize the error measure to 42.27%, the trend was practically constant.

Alfa	1
Beta	0.1454882

MAPE	42.27%

In conclusion, this method cannot in fact be the best for trying to predict the amount spent since MAPE, despite being minimized, still presents a very significant value. Furthermore, from the group's perspective, there is not enough concrete information to state that there is relevant and standard consumption behavior over time, as it has always been very unstable over the three years of analysis.

3.2 Predictive Analysis - Restrictive

3.2.1 Linear Regression Models

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,033420458
R Square	0,001116927
Adjusted R Square	-0,003298045
Standard Error	884,4218901
Observations	2044

Observing the regression statistics table, the essential points to take away are the low value of the coefficient of determination, demonstrating that this regression does not predict, at all, the outcome reliably, and the high value of the standard error, thus showing that the estimated coefficient is very distant from the real population coefficient.

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	9	1781640,715	197960,0794	0,3796207	0,94531708
Residual	2037	1593345636	782202,0797		
Total	2046	1595127277			

Moving on to the analysis of the ANOVA table, it is possible to conclude that this regression is, in fact, neither useful nor reliable for predicting Amount_spent since the Significance F is greater than 0.05. As the null hypothesis is not rejected, there is no causal relationship between the independent variables on the dependent variable.

$$Significance\ F = 1 - \frac{SS\ residual}{SS\ Total}$$

Analyzing the formula, SS residuals stands for Residual Sum of Squares, that is, it is the difference between the observed values of Amount_Spent and the predicted values. As the value is extremely high and close to the total value, it shows that the model does not fit the data well.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	1409.57113	83.5668789	16.8675813	7.4259E-60	1245.68567	1573.45658	1245.68567	1573.45658
Female	0	0	65535	#NUM!	0	0	0	0
Male	-2.5925907	39.2728279	-0.0660149	#NUM!	-79.611683	74.4265011	-79.611683	74.4265011
Married	0	0	65535	#NUM!	0	0	0	0
Single	10.8929368	39.7146018	0.2742804	#NUM!	-66.992531	88.7784044	-66.992531	88.7784044
Unemployme	0	0	65535	#NUM!	0	0	0	0
self-employed	80.4462897	76.5404356	1.05102994	#NUM!	-69.659398	230.551977	-69.659398	230.551977
workers	29.6906944	71.7058872	0.41406216	0.67887218	-110.93382	170.315207	-110.93382	170.315207
Employees	66.5562523	70.2816708	0.94699303	0.34375465	-71.275188	204.387693	-71.275188	204.387693
Age	-0.7360161	1.07677166	-0.6835397	0.49434365	-2.8477045	1.37567229	-2.8477045	1.37567229

All variables have a p-value greater than 0.05, which is reflected in their weak relationship. Therefore, without excluding the null hypothesis, the variables present values of 0. Furthermore, and as previously concluded, the standard error presents extremely high values, which reflects a regression that is not very credible. Variables like Male, Female, Single, Married, self-employed and Unemployment have an error in p-value and zero value in coefficient and other aspects of regression analysis since Excel's data analysis tool does not consider them significant or contributory in the regression.

Six of the nine independent variables present this error, demonstrating that this method is not suitable for predicting “Amount Spent” since, despite there being a mere indirect relationship between gender, marital status and employment status, there is no cause-effect relationship strong enough to predict o “Amount Spent”.

In the case of using a Simple Linear Regression, with age being the only independent variable, the regression would still not be able to predict the Amount_spent in a concrete and credible way. Furthermore, the standard error continues to present a similar value compared to that of multiple linear regression.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,015198189
R Square	0,000230985
Adjusted R Square	-0,000258618
Standard Error	883,7300817
Observations	2044

The significance *F* constitutes a value greater than 0.05, thus having the same conclusion drawn in the multiple linear regression studied and analyzed previously. That is, there is no relationship between age and "Amount Spent", making this regression useless for predicting the value of "Amount Spent".

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	368450,4095	368450,4095	0,471780261	0,492247295
Residual	2042	1594758827	780978,8573		
Total	2043	1595127277			

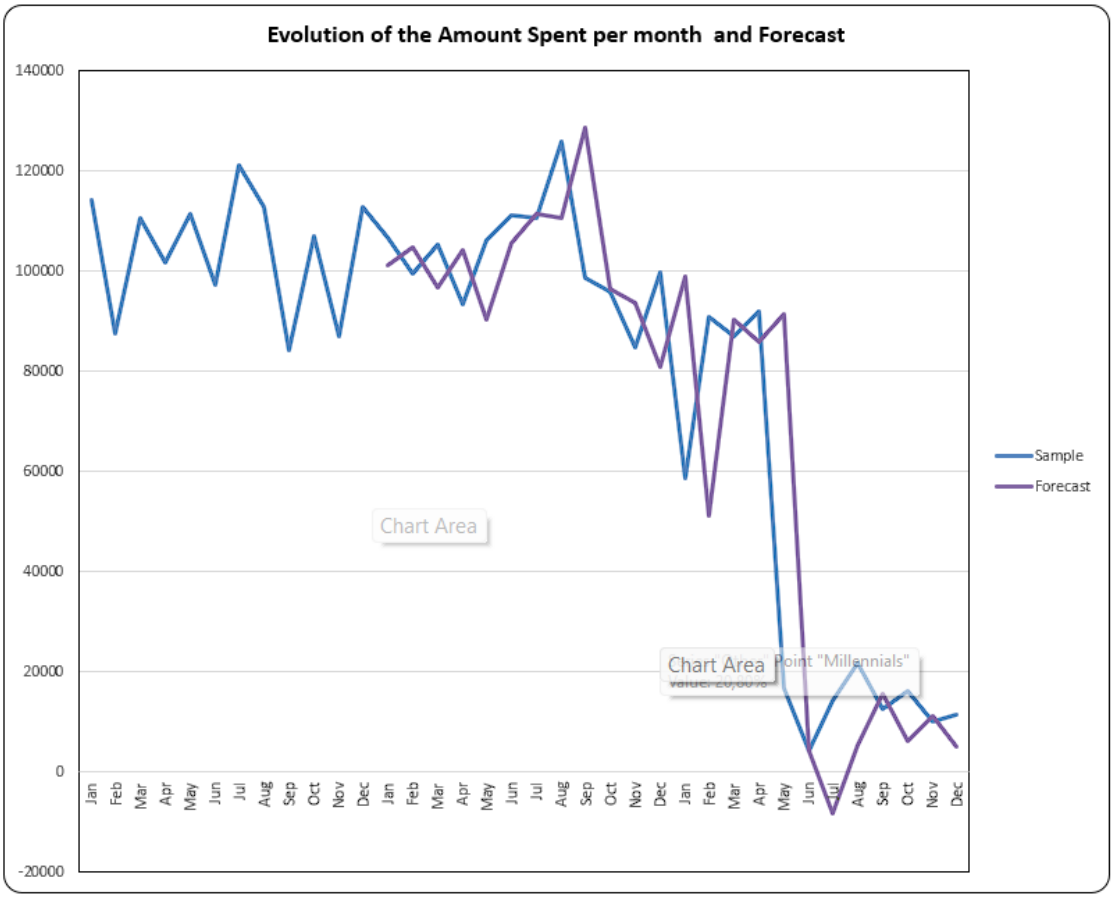
The independent variable age and the ordinate at the origin, separately, have a p-value greater than 0.05, that is, they both have a value equal to zero and do not have a causal relationship with the dependent variable. As in multiple linear regression as in simple linear regression, the only variable that is precise in relation to the proximity between the estimated coefficient and the real population coefficient.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	1464,020153	53,97745044	27,12281037	1,2807E-138	1358,163549	1569,876756	1358,163549	1569,876756
Age	-0,736794495	1,072695574	-0,686862622	0,492247295	-2,840486105	1,366897116	-2,840486105	1,366897116

$$Amount\_Spent = 1464.02 - 0.737 * Age$$

In conclusion, a *Linear Regression Model*, whether simple or multiple, are not good methods for predicting *Amount_Spent* values since there is no strong causal relationship between the independent variables and the dependent variable.

3.2.2 Holt's Linear Trend Method



Starting by observing the coefficient of determination, it presents a better value compared to the linear regression models used, being able to predict the "Amount Spent" in a particular and limited way. As is possible to see from the graph, there is a tendency for the amount spent to decrease over time and there is no seasonality, as there is no repetitive cycle of increase or decrease.

For the analysis and forecast of the total amount spent in the first quarter of 2022, "Transaction Date" was transformed into totals per month. The initialization of the level (St) is based on the average of the 12 months of 2019, and the trend is based on the slope of the trend line. A value of 0.5 was assigned to the alpha and beta in order to discover the level and trend of the real periods of the population, and subsequently calculate the forecast for the months of the years 2019 to 2022.

Together with the calculated information, MAPE was used as an error measure, which refers to the average percentage of error between the values acquired in the forecast and the actual values.

	Forecast
jan	10 694.35 €
fev	11 589.29 €
mar	12 484.23 €

MAPE	85.63%

In order to minimize the error measure, the Solver tool was used, which consisted of minimizing the MAPE by changing the alpha and beta. In order to minimize the error measure, the Solver tool was used, which consisted of minimizing the MAPE by changing the alpha and beta. To minimize the MAPE by half, to 44.80%, the alpha was transformed to 1 and the beta was transformed to approximately 0.16. These values reflect that, to achieve the objective of bringing predicted values as close as possible to actual values, it will be necessary to give more emphasis to current rather than past events, and the trend must be considered constant.

Alfa	1
Beta	0.16423

MAPE	44.80%

In order to complete predictive analysis, both of the methods used to predict the "Amount Spent" are generally not the best methods due to the type of variables and date used. In comparison to linear regression models or the times series forecast model, the one that is best suited is the Holt's linear trend method, despite having a very high error measure, not making it very effective at all.

4. Prescriptive Analysis

Moving on to the third and final analysis, the prescriptive analysis will fall on Moore Pharmaceuticals, where the main objective was to decide whether to go forward to produce the drug after discovering a new drug.

Initially, some data were made available such as the size of the market, revenue and expenditure in units, the discount rate, and the cost of the initial investment. This investment refers to Clinical Trials and R&D expenses, which have a uniform distribution between $800,000 and $600,000. The market size follows a normal distribution with a mean of 2 000 000$ and stand deviation of 400 000$.

Market size		2 101 067
Unit (monthly Rx) revenue	$	130
Unit (monthly Rx) cost	$	40
Discount rate		9%
Project Costs		
R&D	$	749 203
Clinical Trials	$	150 000 000
Total Project Costs	$	150 749 203

$$NORM.INV(RAND(),2000000,400000)$$

$$RAND() * (800000 - 600000) + 600000$$

The model used consists of a profit forecast based on the evolution of revenues and expenses, market growth factor, market size and market share growth rate. The market growth factor follows a triangular distribution of minimum 2%, mode 3% and maximum of 6% and the market share growth rate follows uniform distribution between 25% and 15%.

Year	**1**	**2**	**3**	**4**	**5**	**Average**
Market growth factor		5,00%	5,00%	5,00%	5,00%	
Market size	2 101 067	2 206 120	2 316 426	2 432 247	2 553 860	$ 2 321 944
Market share growth rate		21,37%	19,05%	23,51%	20,04%	
Market share	8,00%	9,71%	11,56%	14,28%	17,14%	
Sales	168 085	214 206	267 774	347 256	437 682	$ 287 001
Annual Revenue	$ 262 213 121	$ 334 161 085	$ 417 726 721	$ 541 719 802	$ 682 784 410	$ 447 721 028
Annual Costs	$ 80 680 960	$ 102 818 795	$ 128 531 299	$ 166 683 016	$ 210 087 511	$ 137 760 316
Profit	$ 181 532 161	$ 231 342 289	$ 289 195 422	$ 375 036 786	$ 472 696 899	$ 309 960 712

$$ROUND(IF(RAND() < = (3\% - 2\%)/(6\% - 2\%),2\% + SQRT(RAND() * (6\% - 2\%) * (3\% - 2\%)),6\% - SQRT((1 - RAND()) * (6\% - 2\%) * (3\% - 2\%))),2)$$

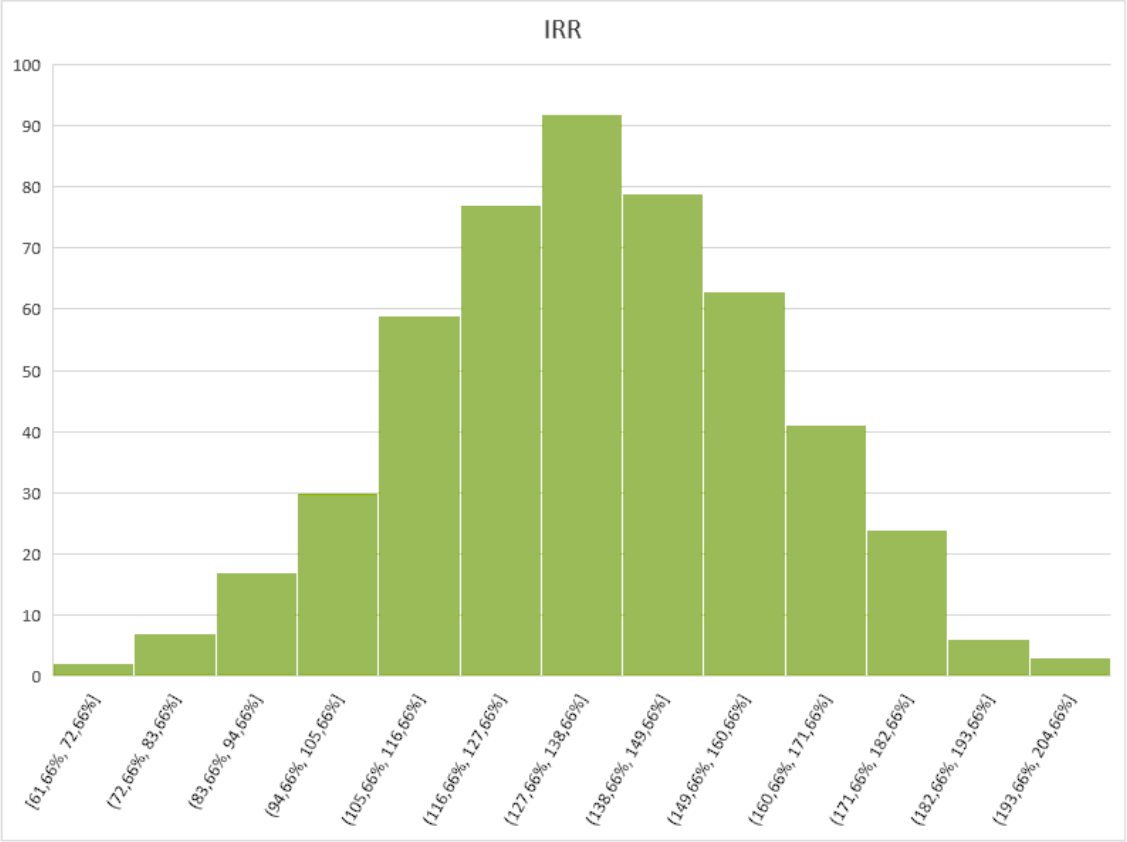
$$RAND() * (0.25 - 0.15) + 0.15$$

The first step was to calculate the Internal Rate of Return (IRR). The IRR is used to evaluate the profitability of an investment. That being said, the cash flow from year 0 to year 5 was calculated and then the IRR was calculated using this information

Years	Profit
0	$ (150 749 203)
1	$ 181 532 161
2	$ 231 342 289
3	$ 289 195 422
4	$ 375 036 786
5	$ 472 696 899

IRR	143%

The cash flow of year 0 is the initial investment and the cash flows of other years are the respective profit.



As it's possible to see, the IRR follows a normal distribution, and the highest incidence of data is between 107.86% and 162.86%. According to these results, it is expected that the value of the investment will more than multiply at the end of the investment period, being quite attractive.

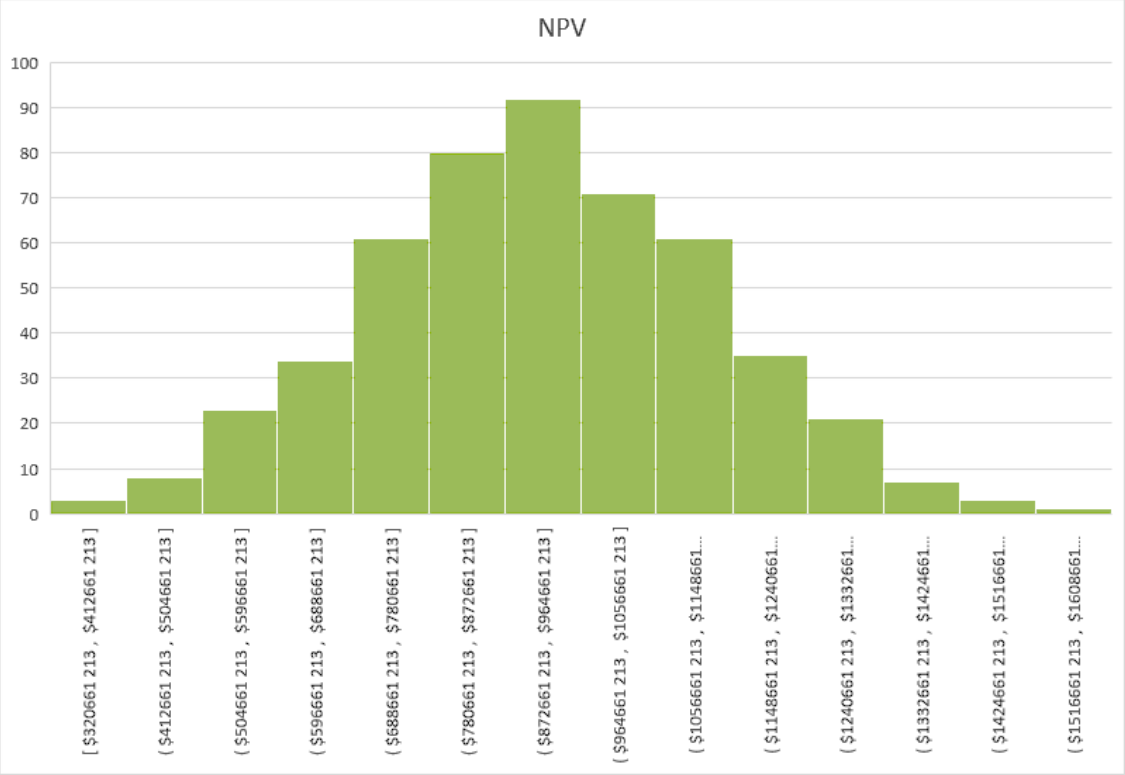
The second step was to calculate the Net Present Value. This value is obtained by the sum of all cash flows divided by one plus the discount rate increased by the number of periods

$$NPV = \sum_{t=0}^{n} \frac{CFt}{(1+r)^t}$$

Years	Yearly NPV
0	$ (150 749 203)
1	$ 166 543 267
2	$ 194 716 177
3	$ 223 311 927
4	$ 265 685 514
5	$ 307 220 551

NPV	$ 1 006 728 233

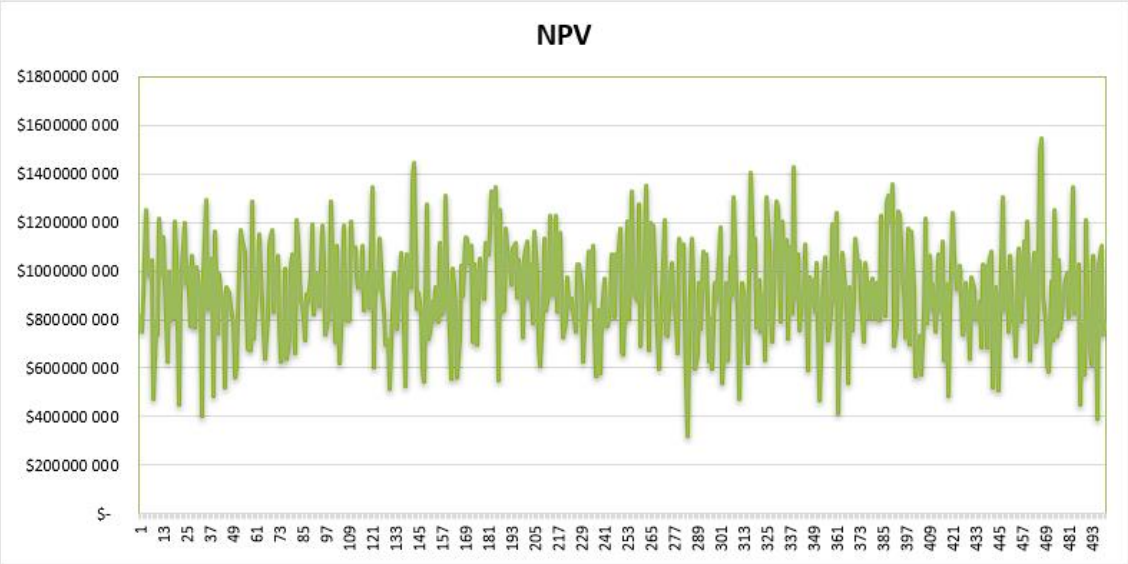
This value represents that the initial investment of approximately 150 000 000$ is expected to generate a cumulative cash flow of $1,000,000,000 over the five-year period. Like the IRR, the NPV follows a normal distribution, in which there is a greater incidence of data between $621,434,619 and $1,111,434,619. Thus, in general, the business appears to be attractive.



The third and final step is the simulation of 500 hypotheses of the values of IRR, NPV, average market size, average annual revenue and average annual costs. In the simulation, the probability of the NPV being greater than $800,000,000 was removed, that indicates that the investment is projected to generate a net return of 5 times higher the initial investment after discounting all future cash flows to present value. This probability is greater than 70%, so a return of 5 times more in current terms can be expected. The probability of the IRR being greater than 100%, that is, the rate of return on the investment annually, was also calculated. The probability is more than. 90% therefore the investment is projected to double the value of the initial investment annually.

Nº de observações	500
P(NPC ≤ 0)	0
P(NPV ≥ 800.000.000)	72.60%
P(IRR ≥ 100%)	91.20%

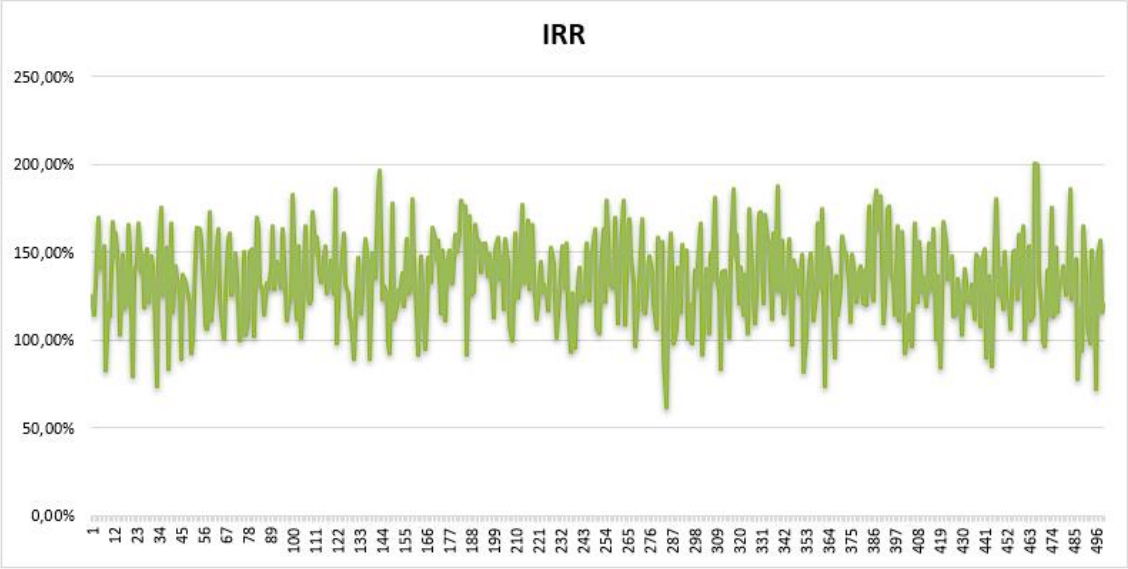
The NPV over the 500 observations always presented values higher than the initial investment.



The greatest probability of the NPV assuming values is between $900,000,000 and $1,200,000,000.

NPV			%	Total	Number
Up to	$	300 000 000,00	0,00%	0	0
Up to	$	600 000 000,00	7,20%	36	36
Up to	$	900 000 000,00	40,00%	236	200
Up to	$	1 200 000 000,00	43,20%	452	216
Starting from	$	1 200 000 000,00	9,60%	48	48
Total			100,00%		500

In turn, the IRR also always presented values above 50%, therefore a return of at least 50% on the initial investment is expected.



The highest probability of values that the IRR could assume is between 130% and 160%. Therefore, the initial investment more than multiplies

IRR		%	Total	Número
Up to	100,00%	8,00%	40	40
Up to	130,00%	34,20%	211	171
Up to	160,00%	42,60%	424	213
Up to	190,00%	14,60%	497	73
Starting from	190,00%	0,60%	3	3
Total		100,00%		500

The average revenue, expenses and profit annually were also analyzed. In the 500 observations, there was an average profit of approximately $280,000,000, hoping for good signs in the future.

Aver. Annual Revenue	$	410 891 977
Aver. Annual Costs	$	126 428 301
Aver. Annual Profit	$	284 463 677
Aver. Market Size	$	2 184 042

In conclusion, after the prescriptive analysis with a study of a simulation of 500 observations, the values obtained in relation to IRR, NPV and profit, demonstrate that the

company Moore Pharmaceuticals must continue with the investment in the launch and development of the new drug, since everything indicates that you could have an extremely appealing and positive return.

5. Conclusion

Throughout the lifespan of the assignment given, from the cleansing and organization of the base datasets, all the way to the handling and analysis of the results, the theoretical-practical direction taken by the group enabled a plethora of conclusions and takeaways. The methodology and mindset behind the drawing of 2 similar, yet distinct, scenarios sprouted from a small academic seed which was only able to bloom due to the tools, cases and knowledge absorbed during the Business Analytics course.

The main conclusion that was drawn lies in the Descriptive and Predictive Analysis which translates in the fact that, regardless of the model being Logical or Restrictive, none of the methods learned and applied were optimal to predict future spending, due to both the models' specific behaviors and limitations, as well as the quality of the data itself.

Regarding the Descriptive Component, the group's aim to identify potential stronger variables and categories in comparison to others, via the removal of observations, and its hypothetical impact on results shown, was proven not to be completely met. In spite of this, the existing respective adjustments and distinctions made from one to model to another enabled the group to branch out from their initial standpoint and opt to provide a holistic overview of the task at hand.

On the Predictive component, the previous diagnosis derives from the quintessential fact that due to the massively heavy, and somewhat recent, observations presented from the dataset, which drastically diverge from the 2 initial years, it does not provide the necessary informational tools and data to sustain a credible foresight. In addition, since there is no strong causal relationship between the independent variables and the dependent variable in the regression models, predicting future values of amounts being spent proves to be non-ideal.

Despite the previous mentions, the elaboration of this report highly promoted critical thinking by striving to find efficient solutions and proxies for the challenges and setbacks the group faced itself with. In fact, the belief that working around complex obstacles through proper application of digital tools, correct interpretation of information and a 'keen eye' is the cornerstone of Business Analytics, and by extension, proof of its imperative value.