TopUpMAMA REPORT

Andrea Giuliodori

February 21, 2022

The provided data covers the period between January 1st, 2022 and February 17th, 2022. The data includes information about the customers, orders and deliveries made by TopUpMama stores in Kenya and Nigeria.

1- Data Cleaning and Merging

Through the six provided databasets, I initially joined them into three, joining the information of Kenya and Nigeria. The cleaning process was necessary previously to this merge. I also added a new variable called "Country" in order to know the origin of the transaction. Then, I obtained the following databases:

Customers: It is composed with the customers of both countries, obtaining 5272 entries

Orders: It is composed with orders of both countries, obtaining 13671 entries.

Deliveries: It is composed by the deliveries of both countries, obtaining 530002 entries.

With the three databases information I merge them into one, by considering the customer ID to merge the "*Customers"* and "*Orders"*. Then, I consider, the Order ID to merge everything with the "*Deliveries"* set. This whole dataset is called *Database_complete* and contains around 13900 entries, one per order and delivery. Some entries of "*Deliveries*" had to be omitted because they do not have the corresponding information in the "*Orders"* set.

All cleaning and merging process were done with Rstudio and Python (Jupyter Notebook), as can be seen in the files attached.

According with the information obtained in the final database, I dropped those variables that were empty or that have the same output for every entry. These type of variables do not have variability, so they do not provide discriminant information.

2- EDA (Exploratory Data Analysis) for all data

The first analysis consist on explore the whole database. The following analysis was performed.

2.1. Quantity of orders by Product Category (Category Name)

In Figure 1, we can observe that the category most demanded is Flour& Sugar.

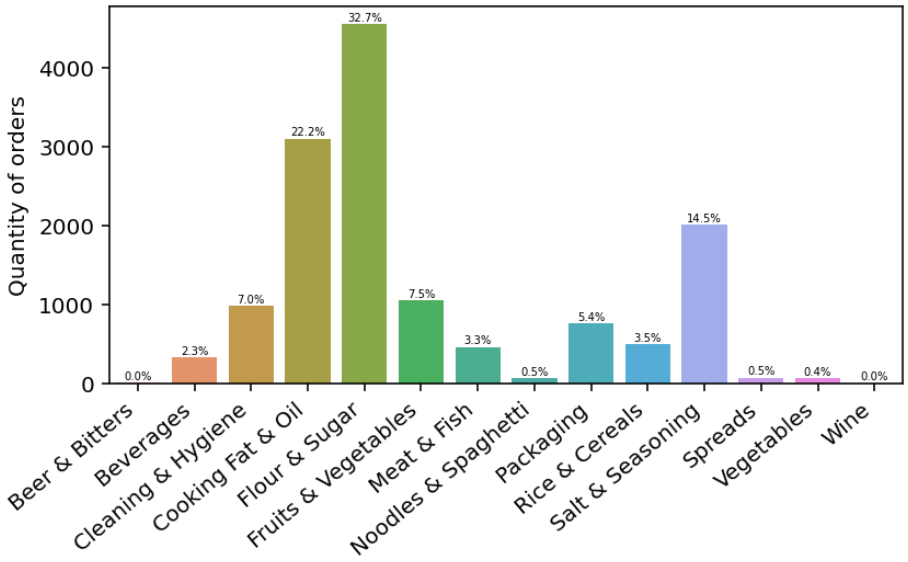


Figure 1: Quantity of orders by product category

If we analyze the category of products by Country, in Figure 2, we confirmed that Flour& Sugar is also the most demanded category of product in both countries. However, Cooking Fat & Oil is also too demanded in Kenya, but not in Nigeria.

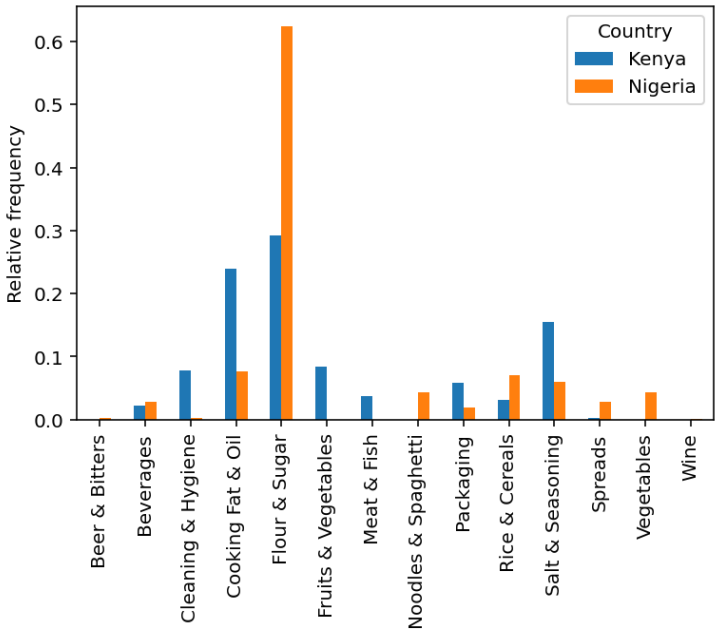


Figure 2: Product Category by Country

According to this data, we can analyze the relationship between the Country and the Category of product that customers demand. This relationship could be studied with the Statistical Test for qualitative variables called Chi-square. The results of the test are shown in Figure 3.

	Chi-square test	results
0	Pearson Chi-square (13.0) =	2349.3647
1	p-value =	0.0000
2	Cramer's V =	0.4105

Figure 3: Chi-square Test- Country&Category Name

The Cramer's V factor gives an idea of the intensity of the relationship. According to Haldun Akoglu , we can interpret it as following:

Interpretation of Phi and Cramer's V.

Phi and Cramer's V	Interpretation
>0.25	Very strong
>0.15	Strong
>0.10	Moderate
>0.05	Weak
>0	No or very weak

Considering the country and the category name, the factor is equals to 0.4105, which shows a very strong relationship among the variables. If this factor is greater or equals to 0.25, the relationship between the variables is very strong. In this case, that results could implied that the product category demanded depends on the Country that the order is done. The needs of the customers are different according to the country. That conclusion could leads to have different sales strategies in Kenya than in Nigeria.

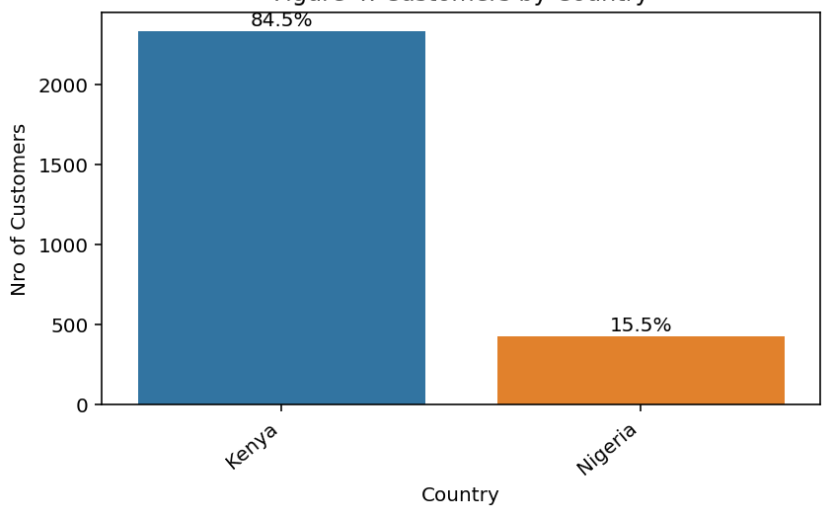
3. Data grouping

I performed a second analysis that consists on grouping the data by customers, by order Id and by stores, in order to have an idea of the behavior of each elements of the process.

3.1. Group by Customer ID

The data were group by Customers, obtaining a database of 2811 customers of both countries. The customers from Kenya represent the 84.5% and the 15.5% are from Nigeria.

Figure 4: Customers by Country

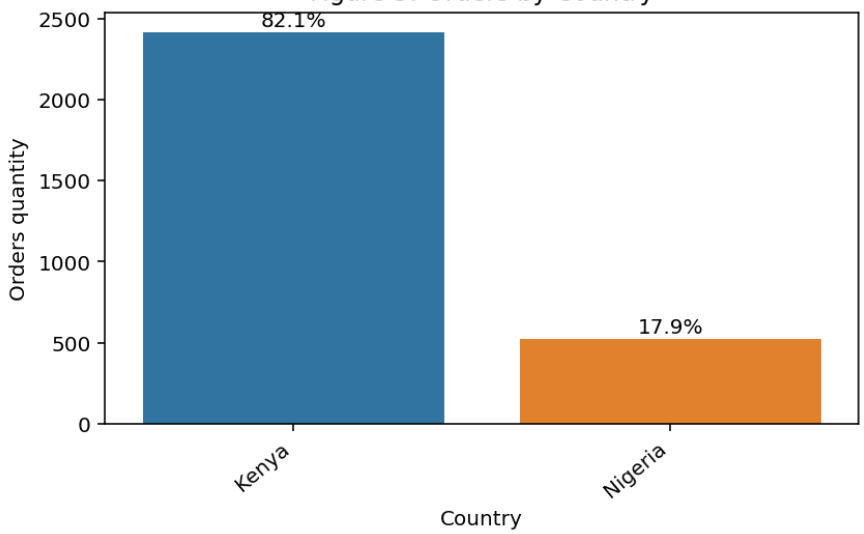


3.2. Group by Order ID

The data were group by Order ID, in an attemp to analyze the characteristics of the orders.

a- Orders by Country

Figure 5: Orders by Country



In Figure 5 we can observe that the 82,1% of all orders come from Kenya, whereas the 17,9% come from Nigeria.

b- Task Status by Country

In Figure 6 we can see the status of the orders depending on the country. We can observed that Nigeria has more *Failed orders* than Kenya.

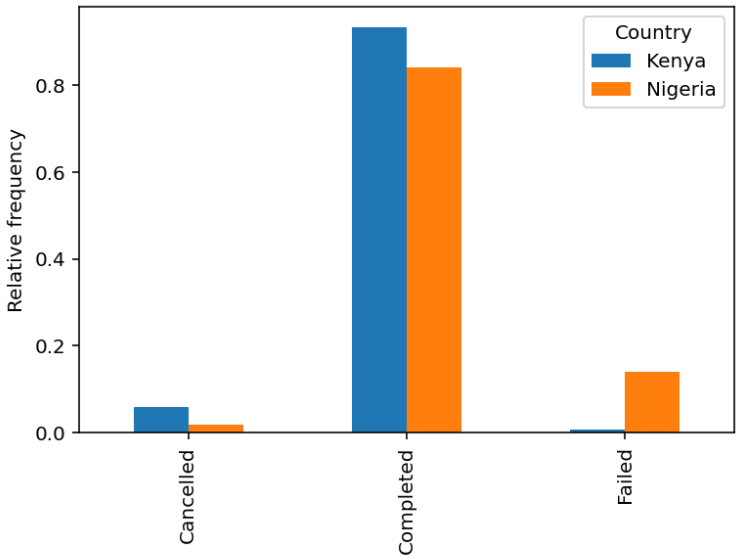


Figure 6: Task Status by Country

In order to know if there is relationship between the origin country of an order and the status of this order, I performed a Chi-Square test. The Cramer`s V factor is 0.2988. That result shows a very strong relationship between variables. Then, some actions can be done depending on the country, for example, to avoid the great number of failed orders in Nigeria.

	Chi-square test	results
0	Pearson Chi-square (2.0) =	262.7955
1	p-value =	0.0000
2	Cramer's V =	0.2988

Figure 7: Chi-square Test- Country&Task Status

c- Payment method by Country

If we analyze the payment Method of an order by country, we can observed that in Kenya there are orders that are *payed later*. Nevertheless, in Nigeria, this characteristic is not observed. The Pay Later of an order is more risky for TopUpMama because it could be transformed in uncollectible.

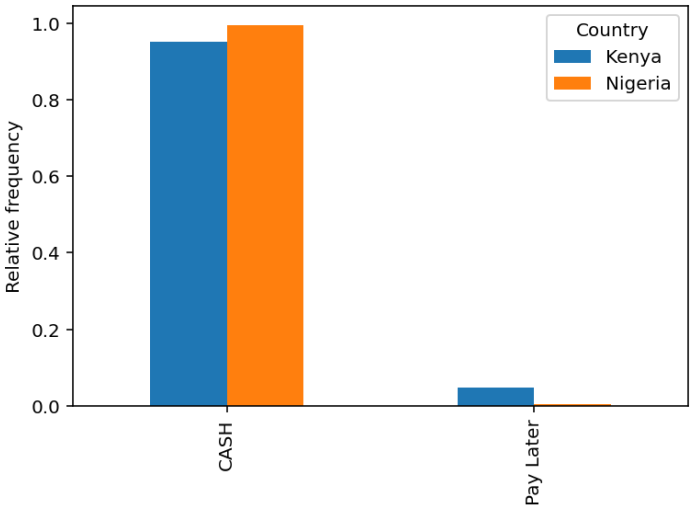


Figure 8: Payment Method by Country

If we perform the Chi-square Test, we can see that there is not a relationship between the variables Country and Payment Method because the Cramer`s Phi factor is too low. Although there is a group of customers that pay later, this fact does not imply the relationship between the variables.

	Chi-square test	results
0	Pearson Chi-square (1.0) =	22.1516
1	p-value =	0.0000
2	Cramer's phi =	0.0868

Figure 9: Chi-square Test- Country&PaymentMethod

d- Order Status by Country

The status of the orders by countries does not show an evidence of relationship among variables.

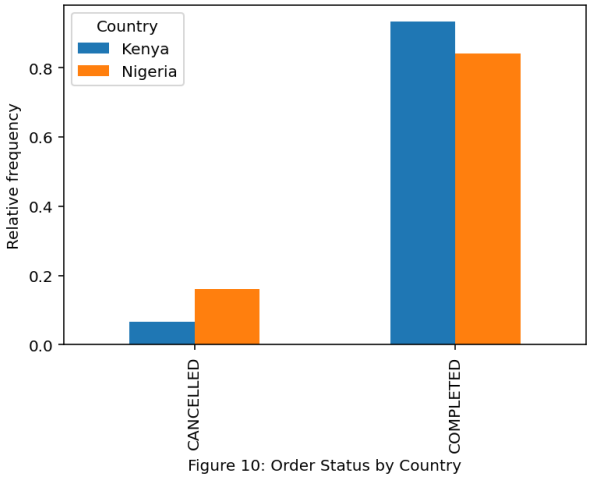


Figure 10: Order Status by Country

This fact is confirmed by the Chi-square Test, where the relationship between variables seems to be moderated.

	Chi-square test	results
0	Pearson Chi-square (1.0) =	48.4307
1	p-value =	0.0000
2	Cramer's phi =	0.1283

Figure 9: Chi-square Test- Country&Order Status

3. Prediction with Python

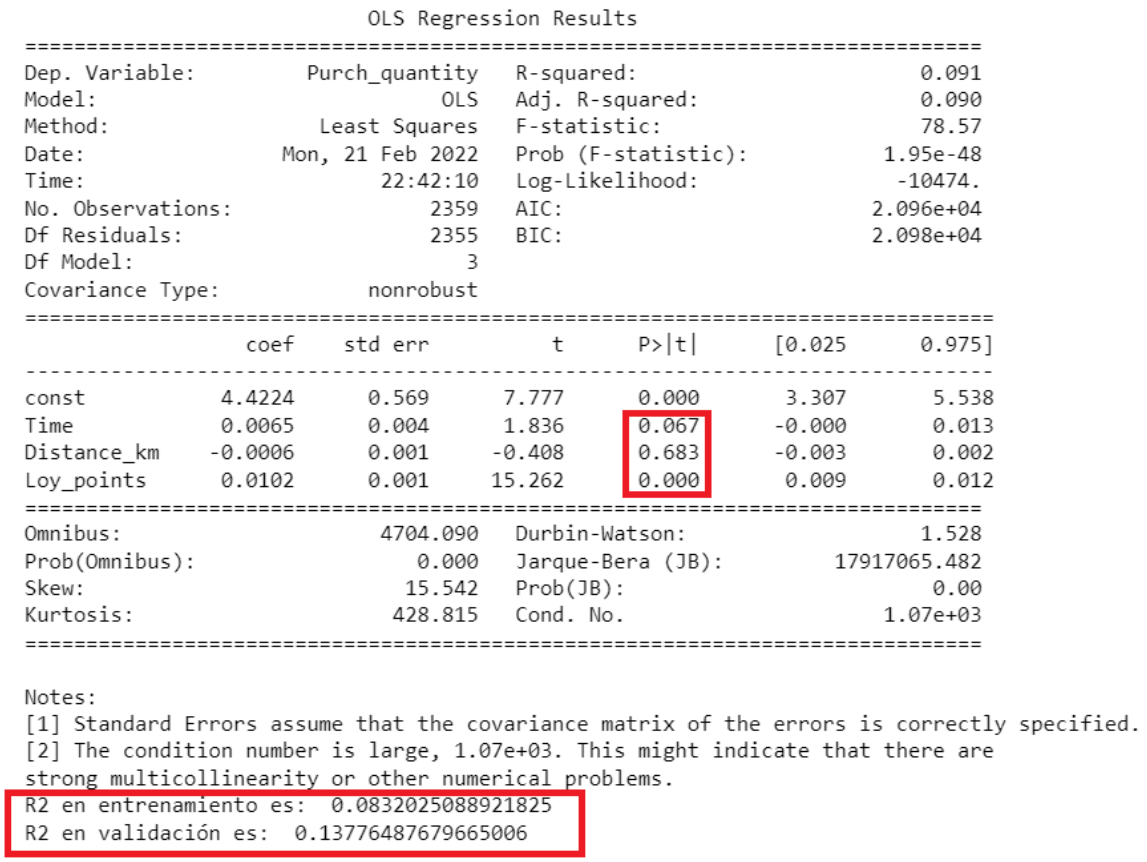
In order to do some predictions I will work only with Kenya orders and customers because there are some inconsistencies in Nigeria data that could affect the results.

3.1. Prediction of the Customers purchase quantity (Using the Database grouped by Customers)

The goal of this modelling is to predict the quantity that a store could order (Purchase quantity), according to the provided data. This variable was obtained from the grouped data by store.

The predictors variables are the *Total time taken*, the *Distance in km* and the *Loyalty Points*.

The results of the regression analysis are shown in Figure 10. According to the output, the *Time* and *Loyalty Points* are significant variables to predict the quantity of a purchase. The *Distance in km* seems not to be significant to predict the quantity of a purchase. However, due to the low value that the R-square has in the model, others variables should be considered in futures studies, in order to increase the amount of variability explained by the model.



OLS Regression Results

Dep. Variable:	Purch_quantity	R-squared:	0.091
Model:	OLS	Adj. R-squared:	0.090
Method:	Least Squares	F-statistic:	78.57
Date:	Mon, 21 Feb 2022	Prob (F-statistic):	1.95e-48
Time:	22:42:10	Log-Likelihood:	-10474.
No. Observations:	2359	AIC:	2.096e+04
Df Residuals:	2355	BIC:	2.098e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.4224	0.569	7.777	0.000	3.307	5.538
Time	0.0065	0.004	1.836	0.067	-0.000	0.013
Distance_km	-0.0006	0.001	-0.408	0.683	-0.003	0.002
Loy_points	0.0102	0.001	15.262	0.000	0.009	0.012

Omnibus:	4704.090	Durbin-Watson:	1.528
Prob(Omnibus):	0.000	Jarque-Bera (JB):	17917065.482
Skew:	15.542	Prob(JB):	0.00
Kurtosis:	428.815	Cond. No.	1.07e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.07e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

R2 en entrenamiento es: 0.0832025088921825
R2 en validación es: 0.13776487679665006

Figure 10: Regression analysis

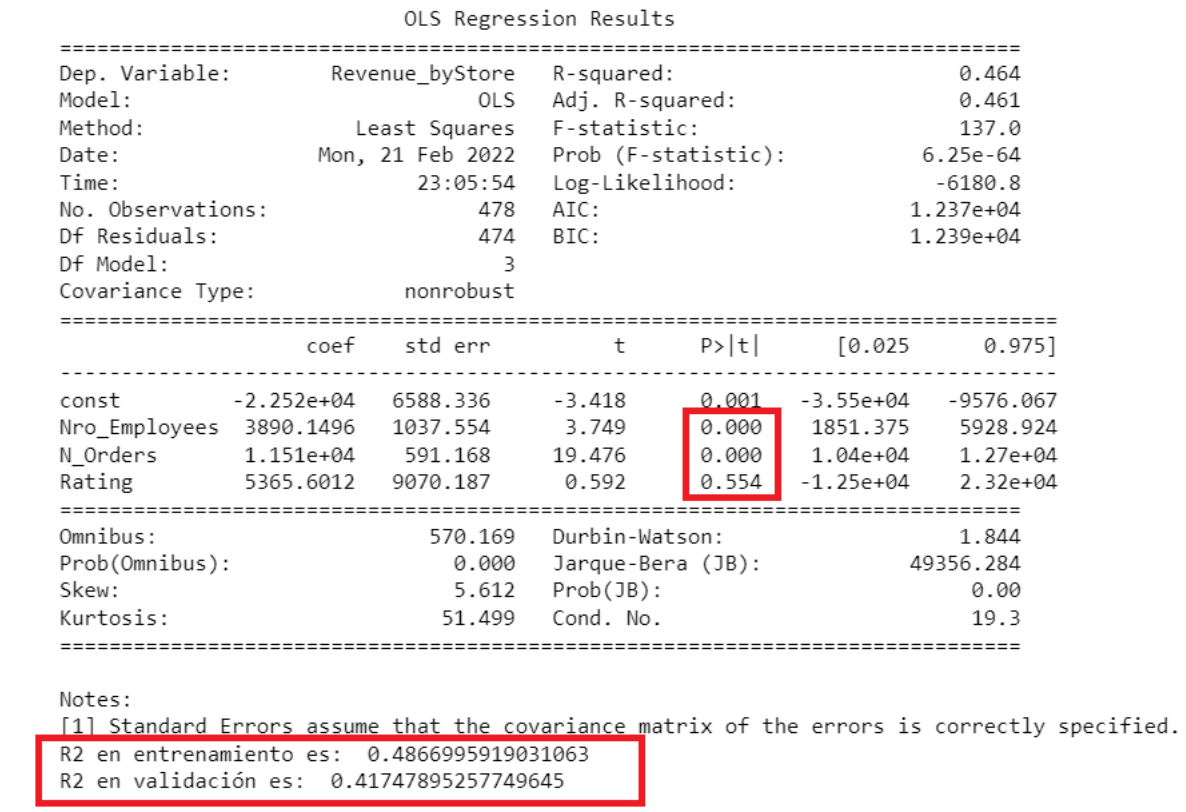
3.2. Prediction of the revenue by store (Using the Database groupby Stores)

In this case, we want to predict the revenue obtained by store that a customer will order. The predictors variables are, the number of employees that the store has, the numbers of orders that the customers perform, and the rating of the delivery.

According to the regression model, the number of employees and the numbers of orders are significant to predict the revenue obtained by store.

The Rsquare of the model is greater than the previous model. However, the inclusion of others variables should be considered in order to achieve a more precise predictor model.

Considering the lack of information provided to carry on the models, It is important to remark that all analysis performed in this report are provisional an subject to review. Further and deeper studies should be done to have an accurate conclusion about all facts analyzed in this report.



OLS Regression Results

Dep. Variable:	Revenue_byStore	R-squared:	0.464
Model:	OLS	Adj. R-squared:	0.461
Method:	Least Squares	F-statistic:	137.0
Date:	Mon, 21 Feb 2022	Prob (F-statistic):	6.25e-64
Time:	23:05:54	Log-Likelihood:	-6180.8
No. Observations:	478	AIC:	1.237e+04
Df Residuals:	474	BIC:	1.239e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2.252e+04	6588.336	-3.418	0.001	-3.55e+04	-9576.067
Nro_Employees	3890.1496	1037.554	3.749	0.000	1851.375	5928.924
N_Orders	1.151e+04	591.168	19.476	0.000	1.04e+04	1.27e+04
Rating	5365.6012	9070.187	0.592	0.554	-1.25e+04	2.32e+04

Omnibus:	570.169	Durbin-Watson:	1.844
Prob(Omnibus):	0.000	Jarque-Bera (JB):	49356.284
Skew:	5.612	Prob(JB):	0.00
Kurtosis:	51.499	Cond. No.	19.3

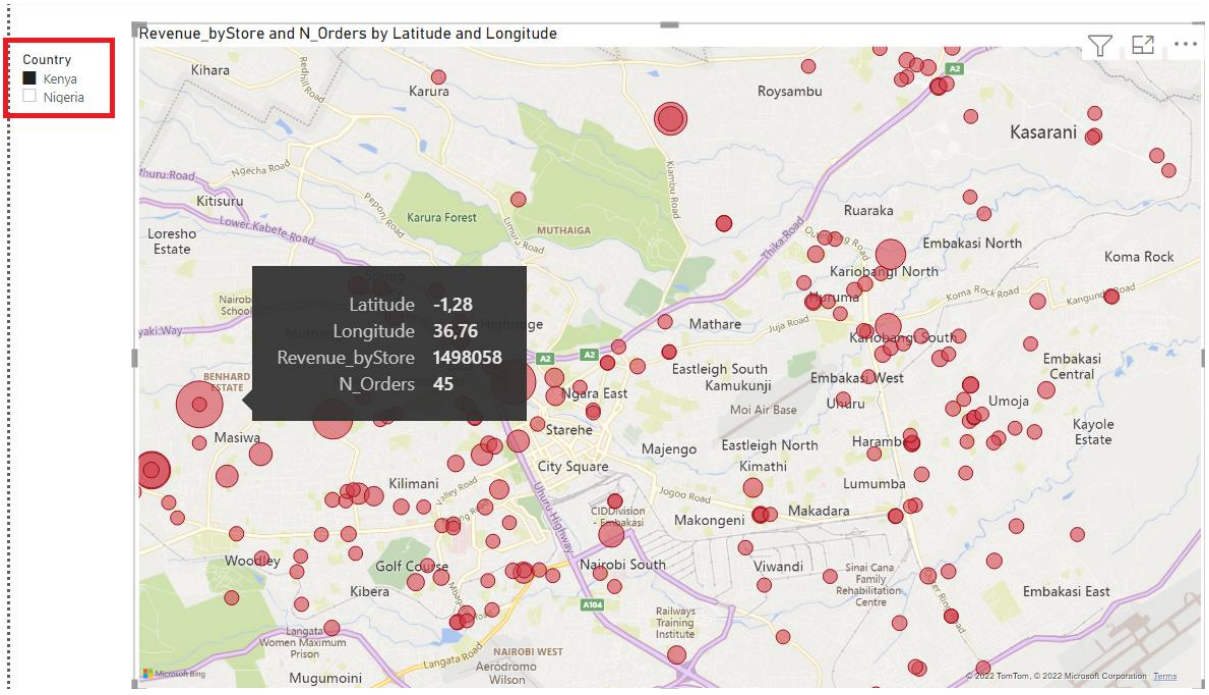
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

R2 en entrenamiento es: 0.4866995919031063
R2 en validación es: 0.41747895257749645

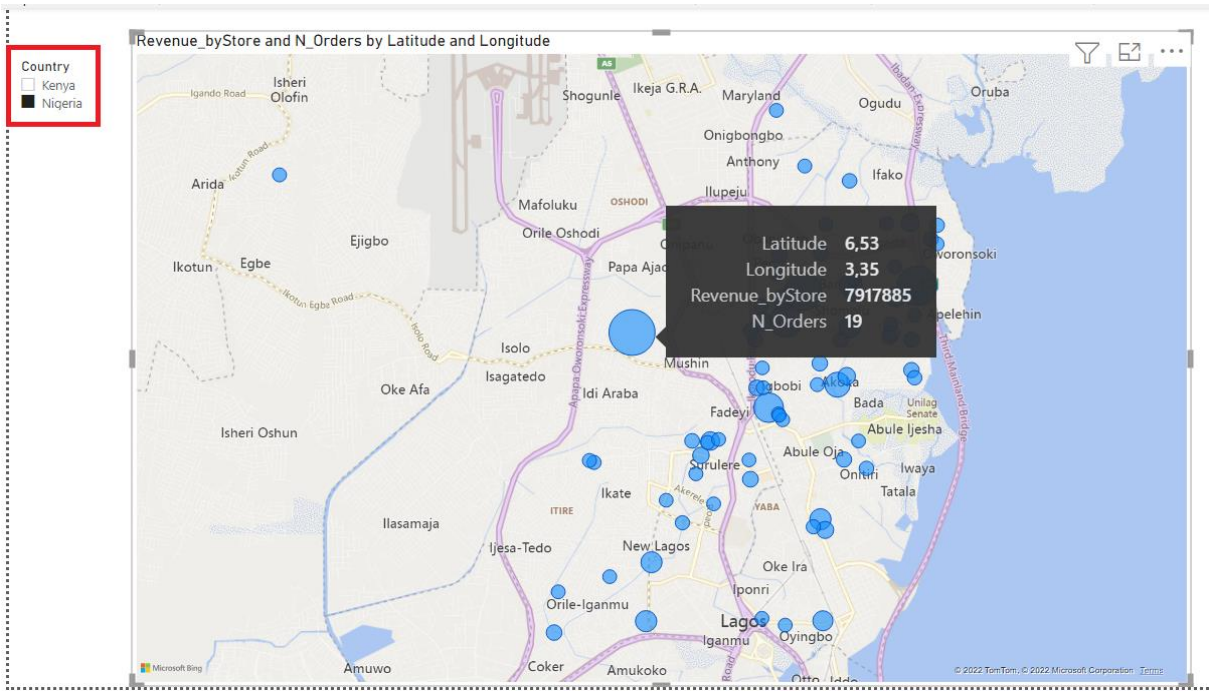
Figure 11: Regression analysis

4. Geospatial analysis with PowerBI

In order to analyze the performance of the stores, we can use the PowerBi software to draw in the map the location of the stores, for example, in Kenya. We can assigned a circle with different size according to some characteristic. In this case, I choose the RevenueBystore in the representation of the map. As an example, the store located in -1,28 latitude and 36,78 longitude has 45 orders and TopUpMama obtains by this store a revenue equals to 1,498,058 KSh in the analyzed period.



An example in Nigeria is the following



Additional interesting analysis can be done in the future. The PowerBi dashboards are dynamics, and they can be customized according to the goals. Besides, those reports can be updated immediately after a data is loaded and shared with all

5. Conclusions and Recommendations:

- It might be important to do the tracking on an order, to avoid great number of failed.
- The Nigerian market is smaller than the Kenyan market. Considering the relationship between some variables analyzed in this reports and the country of origin of an order, certain strategies could be carried out to expand the market in Nigeria.
- The time it takes for a delivery seems to be significant in predicting the quantity of a purchase.
- The number of employees in a store seems to be significant in predicting TopUpMama's revenue.
- Further studies could be done on the great performance of some stores. It might be interesting to explore the reasons for the buyback, the agent in charge, etc.
- Nigerian data should be more precise to allow comparison with Kenyan data.
- The description of each variable could contribute to the interpretation of the analysis performed in this report
- Some variables have a significant number of missing values. It is advisable to be aware of the importance of accurate information to make good estimates.
- In the event that the data set contains a reasonable number of missing values, an estimate of those values can be made. There are some methods to do this imputation, as an example, the Random Forest algorithm could be applied.