DATA ANALYSIS REPORT

TRANSACTIONS ANALYSIS OF AN ONLINE SHOP  
LOCATED IN UNITED KINGDOM

Prepared by Baakh Nikolay

Table of contents

1. Executive Summary

This place will have details at the end of the research

1. Introduction

That dataset is containing transactions of an online shop in UK. Data provided by the online school “Netology”

1. Acronyms

LTV Lifetime Value

ARPC Average Revenue Per User

APPC Average Purchase Per Customer

RPR Repeat Purchase Rate

JN Jupyter Notebook

UK United Kingdom

1. Background for the report

That report was created with data provided by the online school “Netology”. Task was different from created report, but author decided to make in more detailed and interesting for his future employer. In this data research Python JN and Excel has been used as the most useful tools for the objectives set.

Report is written for several different audiences at the same time:

Primary audience:

A primary client, or collaborator. At first reads the introduction and perhaps the conclusion to find out what author did and what his conclusions were.

Later fishing through the body of a report and stopping only for additional details. Therefore, we should provide the main visual evidence from our analysis in the Body, but save more detailed evidence, and other ancillary material, for the Appendix.

Secondary Audience:

* 1. An executive person.

Probably only skims the executive summary and perhaps the conclusion. Therefore, we should leave some type of “headers” in each part to make it easy for this person to dive in, find the “headlines” of our work and conclusions, and close a report.

* 1. A technical supervisor.

Reads the Body and then examines the appendix for quality control. How good a job did we do in raising and answering the interesting questions? How efficient we were? Did we reach reasonable conclusions? Used correct statistical methods? Etc.

Here we should make cross-references between the body and parts of the appendix, so that this person can easily find supporting material related to each main analysis we report in the Body.

Therefore, all metrics used are made to be as detailed as possible.

1. Objectives

Data review. Understanding and explaining dataset. Their integrity, quality and anomalies

Data research. Detailed explanation of a work with dataset. Limitations, interpretations, assumptions that had to be done.

Data cleaning.

Business research. Metrics. Explanation of a business client (UK online shop provided its data). Its pain points and targets.

Recommendations. Data structure recommendation. What to add to the dataset. Formulate conclusions and recommendations based on the data provided.

1. Body

Data

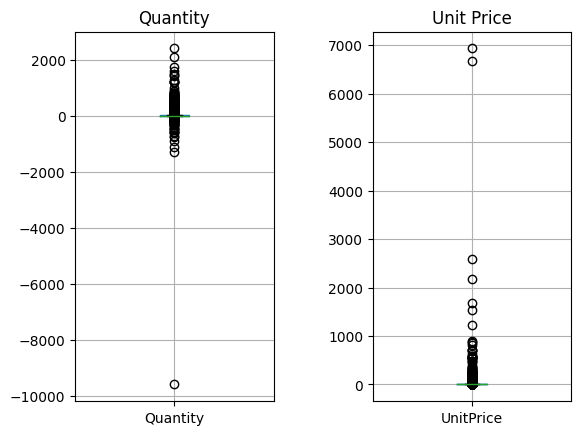
We have dataset “Data Sample.xlsx” with 54214 rows

|  |  |  |  |
| --- | --- | --- | --- |
| № | Column | Non-Null Count | Data type |
| 0 | InvoiceNo | 54214 non-null | Object |
| 1 | StockCode | 54214 non-null | Object |
| 2 | Description | 54080 non-null | Object |
| 3 | Quantity | 54214 non-null | Int64 |
| 4 | InvoiceDate | 54214 non-null | Datetime64[ns] |
| 5 | UnitPrice | 54214 non-null | Float64 |
| 6 | CustomerID | 40643 non-null | Float64 |
| 7 | Country | 54214 non-null | Object |

Dataset had no detailed documentation so we will make assumptions about each column based on research work:

* “InvoiceNo” – Number of invoices
* “StockCode” – Code of an item
* “Description” – Description of an item
* “Quantity” – Sold quantity of that exact item in that invoice
* “InvoiceDate” – Date of an invoice
* “UnitPrice” – Price of a bought unit
* “CustomerID” – ID of a customer
* “Country” – Country of a buyer

As we can see in the table, presented data have some problems. Some values are missing. That way we are having "null" values in "Description" and "CustomerID" columns.



Also, we have some anomalies. They are shown on a boxplot. Negative and too high values in “Quantity” column. Some values in “UnitPrice” column are equal to 0 or too big.

That data has a unique identifier as a combination of “InvoiceNo’ and “StockCode”. But data has duplicates, that must be deleted.

Data has 1083 rows with negative Quantity. All of them using “C” at the beginning of every value in “InvoiceNo”. I created an assumption that this is a return of a product.

225 rows had price equal 0. Each of it had something weird in the description, so we’ll dele this from our data sample.

Problem with “null” values in description depends on an “Invoice” column, not on a “StockCode” column as research shows.

One invoice can have only 1 “CustomerID” as research shows.

Data Cleaning

I have decided to delete highest anomalies, rows price of which equals 0, that has no description or CustomerID. Also, duplicates were deleted based on a combination of an “InvoiceNo” and “StockCode” values.

Business research

We found all metrics that we were possible to from that database

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Amount | Description | Who can use that |
| Average Price | 3.3 | Average price of a unit in our company |  |
| APPC | 68 | Average revenue per 1 purchase (invoice) of a customer |  |
| Unique Customers | 3819 | Amount of our unique customers in the lifetime of a dataset |  |
| ARPC | 232 | Average revenue per 1 customer in the lifetime of a dataset |  |
| Percentage Of Returns | 2.26 % | Percentage of returned goods in the lifetime of a dataset |  |
| RPR | 59.23 % | Percentage of users, that made their second purchase in the lifetime of a dataset |  |
| Gross Profit | 916647.11 | Whole profit of a company minus losses on a return |  |
| LTV | 493.133 | Expected lifetime value of a customer |  |

We are compared some data in search for dependencies and trends

Chart, line chart

Description automatically generated

Here we can see that our company grows in good pace, but we have to check that in a longer term, because it can be seasonal.

Chart, bar chart

Description automatically generated

That graph shows us our main countries to work with. Of course, our main customers are in the United Kingdom, but after that we can see EIRE, Germany, and France, they are main of the leading countries in both graphs.

1. Discussion
2. Conclusion
3. Recommendations
4. Appendices