Example Manuscript Template for a Data Analysis Project

Andreas Handel

2024-03-07

# 1. Summary/Abstract

*Write a summary of your project.*

# 2. Introduction

## 2.1 General Background Information

This project aims to gain knowledge of customer segmentation from the supervised and un-supervised learning perspectives. In the initial analysis phase, the study applies basic statistical techniques such as logistic regression to analyze a retail store’s data, to grasp general patterns of customer behavior. As the project progresses, it integrates more advanced tool, Expectation Maximization algorithms.

It is crucial that organizations understand their customer’s behaviors and their purchasing patterns. By studying the purchasing patterns of its customers, an organization can design and execute effective customized customer engagement strategies, which can enhance customer service, and consequently increase sales. Thus, in order to understand their customer’s purchasing behavior, organizations can engage in customer segmentation. This analytical approach involves classifying customers into groups exhibiting similar characteristics or purchasing behaviors. When organizations analyze factors, such as purchase frequency and spending habits, they can identify distinct segments within their customer base and subsequently tailor their marketing and sales efforts to meet the unique needs of each group. This tailored personalized approach can enhance customer satisfaction and also improve multiple end results like customer loyalty, operational efficiency, and profitability.

The project will employ a simple model for classification, logistic regression. Since the data doesn’t have a labeled categorical variable for classification, a target variable will be created based on the purchase amount and frequency during a fixed period.

The project will also explore the use of mixture models as a potential analytical tool based on the assumption that customer data is inherently complex which can exhibit multiple underlying patterns that are not immediately evident through conventional analysis. Thus, the use of finite mixture models is useful in this project. These models are adept at analyzing data that may contain multiple, distinct groups hidden within the data. Gaussian mixture models and mixtures of t-distributions are examples of the most commonly used finite mixture models in analyzing such data. However, models like these that are primarily grounded in symmetric distributions may not always provide an accurate representation of data characterized by asymmetric distributions. Therefore, in cases where skewed data need to be analyzed, asymmetric models become useful since they offer a broader applicability across various model-based clustering scenarios (Lee & McLachlan, 2013). More specifically, the application of a mixture of gamma densities is a rational choice when dealing with observed data that is positive and exhibits right-skewness, suggesting a mixture distribution origin (Young, Chen, Hewage, & Nilo-Poyanco, 2019). This project will also explore the application and compare the results of Gaussian mixture model and Gamma mixture model on the retail store’s data.

## 2.2 Description of data and data source

*Describe what the data is, what it contains, where it is from, etc. Eventually this might be part of a methods section.*

This study employs sales data, ‘superstoredata’ and focuses on international sales of the online retail store.

Initially this data was downloaded from Kaggle.com to assess it’s usability for the project. Upon confirming the usability and revisiting the website seeking for further details, it could not be found in Kaggle.com. After few Google searches, the data and details were found on UC Irvine Machine Learning Repository [https://archive.ics.uci.edu/dataset/352/online+retail]. As per the website, this is a transnational data set containing all the transactions occurring between 12/1/2010 and 12/09/2011 for a UK-based and registered non-store online retail. The company primarily sells unique all-occasion gifts and many customers of the company are wholesalers.

This data has nine variables. The details are explained below:

1. InvoiceNo: It is a categorical variable with a 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter ‘c’, it indicates a cancellation.
2. StockCode: It is a categorical variable with a 5-digit integral number uniquely assigned to each distinct product.
3. Description: This represents Product name and is a categorical variable.
4. Quantity: It represents quantities of each product (item) per transaction and is an integer.
5. InvoiceDate: It represents the day and time each transaction was generated.
6. UnitPrice: It represents product price per unit in sterling.
7. CustomerID: It is a categorical variable with a 5-digit integral number uniquely assigned to each customer.
8. Country: It represents the name of the country where each customer resides.
9. Sales: It represents product price multiplied by UnitPrice.

## 2.3 Questions/Hypotheses to be addressed

*State the research questions you plan to answer with this analysis.*

This project aims to gain a comprehensive understanding of customer segmentation from different perspectives; predictive (logistic regression), probabilistic clustering (GMM), and skewness-aware clustering (Gamma Mixture Models).

# 3. Methods

*Describe your methods. That should describe the data, the cleaning processes, and the analysis approaches. You might want to provide a shorter description here and all the details in the supplement.*

## 3.1 Data import and cleaning

*Write code that reads in the file and cleans it so it’s ready for analysis. Since this will be fairly long code for most datasets, it might be a good idea to have it in one or several R scripts. If that is the case, explain here briefly what kind of cleaning/processing you do, and provide more details and well documented code somewhere (e.g. as supplement in a paper). All materials, including files that contain code, should be commented well so everyone can follow along.*

This project employs sales data , ‘superstoredata’, of an online retail store based in United Kingdom, to explore purchase patterns of customers outside of United Kingdom.

Initially this data was downloaded from Kaggle.com to it’s usability for the project. Upon confirming the usability and revisiting the website seeking for further details, it could not be found in Kaggle.com. After few Google searches, the data and details were found on UC Irvine Machine Learning Repository [https://archive.ics.uci.edu/dataset/352/online+retail]. As per the website, this is a transnational data set containing all the transactions occurring between 12/1/2010 and 12/09/2011 for a UK-based and registered non-store online retail. The company primarily sells unique all-occasion gifts and many customers of the company are wholesalers.

The data has 541909 observations and nine variables detailed as below:

1. InvoiceNo: It is a categorical variable with a 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter ‘c’, it indicates a cancellation.
2. StockCode: It is a categorical variable with a 5-digit integral number uniquely assigned to each distinct product.
3. Description: This represents Product name and is a categorical variable.
4. Quantity: It represents quantities of each product (item) per transaction and is an integer.
5. InvoiceDate: It represents the day and time each transaction was generated.
6. UnitPrice: It represents product price per unit in sterling.
7. CustomerID: It is a categorical variable with a 5-digit integral number uniquely assigned to each customer.
8. Country: It represents the name of the country where each customer resides.
9. Sales: It represents product price multiplied by UnitPrice.

One key factor of this data is it accounts sales of a particular item from an Invoice as an observation leading different items sold in the same Invoice to appear as separate observations. As such most of the Invoice numbers appear several times in different observations of the data.

As the purpose of the project is to study purchase pattern of customers residing outside of UK, first all of the UK based observations were dropped from the data which brought down the data size to 46431 observations with 9 variables.

The data inspection pointed out to several issues requiring cleaning such as missing values in CustomerID, negative signs in Quantity and Sales and incorrect format of date. It also pointed out that InvoiceNo, StockCode, Description and Country were coded as character variables in the data which needed to be changed to factors for better data manipulation.

First, there were 1480 missing values in CustomerID. Customer behavior is a crucial part of this analysis which cannot be tracked without the unique customerID. The observations without CustomerID do not help in doing meaningful analysis and are dropped. Further, CustomerID was coded as a numerical variable in the original data and was converted to factor variable during the cleaning process. Country, Description and StockCode were also converted to factors from character variables. Transaction date was also converted to the date format.

There were 1372 observations with negative values in quantity and sales due to order cancellation which were dropped from the data.

A detailed scrutiny of the distribution of UnitPrice revealed that all the high value goods with a Unitprice above 25 were distributed between 75 and 100 percentile. A closer look at this inter-quartile range disclosed that there were 19 categories of product description which had the Unitprice above 25. Some of those categories had descriptions such as CARRIAGE, POSTAGE, Manual, which appeared to be related to logistics. 1216 observations with such descriptions were dropped from the data since those were less likely to be directly related to customer purchase.

The store also sells merchandises in wholesale quantities. A wholesaler potentially purchased in big volumes and in higher frequencies. In absence of clear information on wholesalers, a scatter plot of Total Invoice Value by Purchase Frequency was plotted to identify potential wholesalers. The scatter plot revealed that most of the customers made purchases less than a frequency of 50 during that period and most of the cumulative purchase values were less than 50,000. Assuming that the outliers represented purchases by wholesalers, observations with purchase frequencies above 50 and cumulative Invoice value above 50,000 are dropped from the data. This step excluded 9745 observations from the data. The cleaned data has 32602 observations with 9 variables.

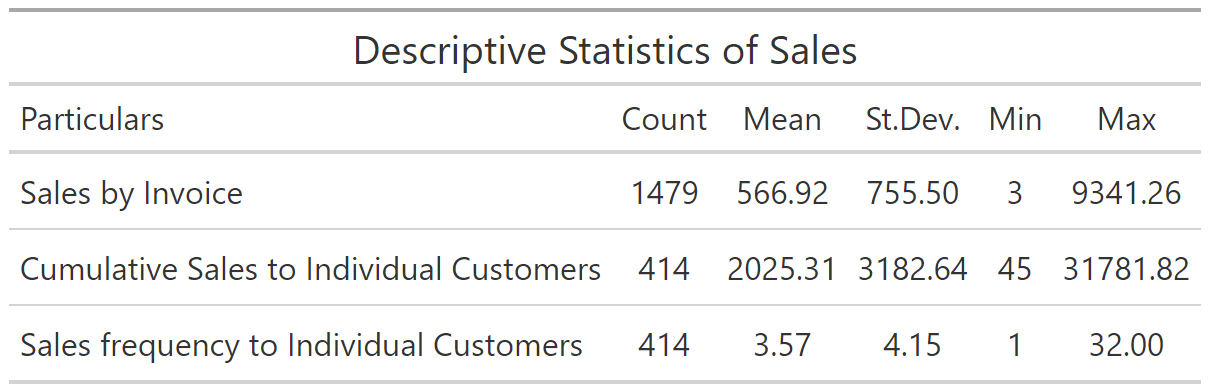
## 3.2 Statistical analysis

*Explain anything related to your statistical analyses.*

# 4. Results

## 4.1 Exploratory/Descriptive analysis

A meaningful analysis of customer spending pattern require the Invoice Sales Amount, total purchase by Customers, frequency of customer purchases during a period or a combination of those. **?@tbl-resulttable1** is created which provides the descriptive statistics of sales by Invoice value, Customer purchase and frequency of customer purchase. It revealed that there were 1479 transactions with 414 different customers during the period of 2010-12-01 to 2011-12-09. The store recorded a mean sales of GBP 567 per transaction, an average cumulative sales of GBP 2025 to individual customers and sales frequency of above 3 times per customer.



**?(caption)**

[Figure 1](#fig-result) depicts two density plots for the Cumulative Sales across customers and Invoice Sales and a histogram for Customer purchase frequencies. Both density plots are heavily right-skewed.The averages of total sales to a customer and total invoice values appear to be less than 2500 and 1250 sterling pounds respectively supporting the mean values of 2025 and 563 in the descriptive table. The average frequency appears to be less than 5 times which is also observed in the descriptive statistics table.

|  |
| --- |
| Figure 1: Distribution of Sales and Purchase frequency |

## 4.2 Basic statistical analysis

*To get some further insight into your data, if reasonable you could compute simple statistics (e.g. simple models with 1 predictor) to look for associations between your outcome(s) and each individual predictor variable. Though note that unless you pre-specified the outcome and main exposure, any “p<0.05 means statistical significance” interpretation is not valid.*

#**?@fig-result1** shows a scatterplot figure produced by one of the R scripts.

## 4.3 Full analysis

*Use one or several suitable statistical/machine learning methods to analyze your data and to produce meaningful figures, tables, etc. This might again be code that is best placed in one or several separate R scripts that need to be well documented. You want the code to produce figures and data ready for display as tables, and save those. Then you load them here.*

#Example **?@tbl-resulttable2** shows a summary of a linear model fit.

**?(caption)**

# 5. Discussion

## 5.1 Summary and Interpretation

*Summarize what you did, what you found and what it means.*

## 5.2 Strengths and Limitations

*Discuss what you perceive as strengths and limitations of your analysis.*

## 5.3 Conclusions

*What are the main take-home messages?*

*Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end*

This paper (Leek & Peng, 2015) discusses types of analyses.

These papers (McKay, Ebell, Billings, et al., 2020; McKay, Ebell, Dale, Shen, & Handel, 2020) are good examples of papers published using a fully reproducible setup similar to the one shown in this template.

Note that this cited reference will show up at the end of the document, the reference formatting is determined by the CSL file specified in the YAML header. Many more style files for almost any journal [are available](https://www.zotero.org/styles). You also specify the location of your bibtex reference file in the YAML. You can call your reference file anything you like, I just used the generic word references.bib but giving it a more descriptive name is probably better.

# 6. References

Lee, S. X., & McLachlan, G. J. (2013). Model-based clustering and classification with non-normal mixture distributions. *Statistical Methods & Applications*, *22*(4), 427–454.

Leek, J. T., & Peng, R. D. (2015). Statistics. What is the question? *Science (New York, N.Y.)*, *347*(6228), 1314–1315. <https://doi.org/10.1126/science.aaa6146>

McKay, B., Ebell, M., Billings, W. Z., Dale, A. P., Shen, Y., & Handel, A. (2020). Associations Between Relative Viral Load at Diagnosis and Influenza A Symptoms and Recovery. *Open Forum Infectious Diseases*, *7*(11), ofaa494. <https://doi.org/10.1093/ofid/ofaa494>

McKay, B., Ebell, M., Dale, A. P., Shen, Y., & Handel, A. (2020). Virulence-mediated infectiousness and activity trade-offs and their impact on transmission potential of influenza patients. *Proceedings. Biological Sciences*, *287*(1927), 20200496. <https://doi.org/10.1098/rspb.2020.0496>

Young, D. S., Chen, X., Hewage, D. C., & Nilo-Poyanco, R. (2019). Finite mixture-of-gamma distributions: Estimation, inference, and model-based clustering. *Advances in Data Analysis and Classification*, *13*, 1053–1082.