Israel Nolazco  
israelnolazco@lewis.edu

Overview of Factor Analysis  
  
  
July 26, 2020

Table of Contents

[I. Background 3](#_Toc46181653)

[II. Executive Summary 4](#_Toc46181654)

[III. Analysis 4](#_Toc46181655)

[a. Designing Factor Analysis 4](#_Toc46181656)

[b. Deriving Factors and Assessing Overall Fit 5](#_Toc46181657)

[c. Interpreting the Factors 6](#_Toc46181658)

[d. Validation of Factor Analysis 7](#_Toc46181659)

[IV. Summary 9](#_Toc46181661)

[APPENDIX I: OVERVIEW OF DATASET 11](#_Toc46181662)

[APPENDIX II: Assessing the Appropriateness of Factor Analysis 11](#_Toc46181663)

[APPENDIX III: Deriving Factors and Assessing Overall Fit 12](#_Toc46181664)

[APPENDIX IV: Validation of Factor Analysis 13](#_Toc46181665)

1. **Background**The purpose of this project is to perform various method of factor analysis in order to further develop a deep understanding in picking variables that correlate with another to create a factor. A factor is made up of multiple variables and its goal is to represent and to narrow the scope of variables being analyzed at once.   
     
   The variables on the HBAT(1).xls dataset are presented in the table below:

* 'x6 Product Quality'
* 'x7 E-commerce'
* 'x8 Technical Support'
* 'x9 Complaint Resolution','x10 Advertising'
* 'x11 Product Line'
* 'x12 Salesforce Image'
* 'x13 Competitive Pricing'
* 'x14 Warranty & Claimns'
* 'x15 Packaging'
* 'x16 Order & Billing'
* 'x17 Price Fexibility'
* 'x18 Delivery Speed'

A screenshot of a cell phone

Description automatically generated

Figure 1. Overview of the dataset and its contents. See Appendix I for code.

Using many of the variables listed above, several visualizations were created that could be used to visually identify variables correlating with another. In a broader scope this could potentially reveal true factors in our data. The next section provides a high-level synopsis of details that are presented in the remainder of this document. Additionally, the information was analyze using the programming language Python.

1. **Executive Summary**  
   This document presents an analysis of factor analysis that were obtained from the HBAT(1).xls dataset.   
     
   Within the aforementioned dataset we are primarily looking at vendor information regarding a product(s). Based on the metrics gather on this dataset, a bigger picture will be painted in order to understand consumer behavior. Given the understanding of what the dataset is representing. We could move to made educated assumption as to variables that may correlate with another i.e Product Quality and Competitive Pricing. Therefore, the type of factor analysis this dataset requires is that of R-type factor and a correlation matrix between its variables.
2. **Analysis**
   1. Designing Factor Analysis

A picture containing track

Description automatically generated

Figure 2. Assessing the Appropriateness of Factor Analysis: Correlations (right side of diagonal divide), Measures of Sampling adequacy and partial correlation among variables(left side of diagonal divide). See Appendix II for code.

As an example of the insight gather from this table, let’s look at variables ('x16 Order & Billing', 'x17 Price Fexibility', and 'x18 Delivery Speed') These variables are inherently correlating with another at a level of 40% to 75%. Therefore, these variables could be considering a factor of their own, i.e a Logistics Factor. However, paying a closer look at columns x15 and x17 we noticed that the correlations are extremely low to be consider, therefore, moving forward the aforementioned variables will be removed.

* 1. Deriving Factors and Assessing Overall Fit

Having identify the importance of 11 variables it now time to derive the factors and start assessing the overall fit of the information gather. In this case we are going to look at the variables with eigen values great than 1. This is to gather meaningful factors out of our variables.

A picture containing drawing

Description automatically generated

Figure 3. Eigen Values. Noticed that only the first four variables are meaning enough based on our criteria.

A close up of a map

Description automatically generated

Figure 4. Eigenvalue vs factor. This graph shows a clear picture where the eigen values are most prevalent in the first four variables. See Apendix III for code.

* 1. Interpreting the Factors

First let us look at the unrotated factor matrices. This represents the degree of associate of each variable with each factor.

A screenshot of a cell phone

Description automatically generated

Figure 5. Unrotated Component Analysis Factor Matrix

Then with a VARIMAX component analysis factor matrices. Varimax is referring to the Orthogonal rotation. The rotation improves the interpretation. Additionally, it is also impacting on the over factor solution and the factor loadings.

A screenshot of a cell phone

Description automatically generated

Figure 6. VARIMAX-rotated Component Analysis Factor Matrices:

* 1. Validation of Factor Analysis

Thus far, our analysis has concluded that we are capable of extracting four factors out of our dataset; however, in order to be cautious, the experiment will be re-run, but this time with a random sample size. In this case, we are going to remove the last forty rows and re-run our analysis.

A screenshot of a cell phone

Description automatically generated

Figure 7. New Eigenvalue vs factor. This graph shows a clear picture where the eigen values are most prevalent in the first four variables. See Apendix IV for code.

A screenshot of a video game

Description automatically generated

Figure 8. New Component Analysis Factor Matrices with Unrotated and VARIMAX rotation.

1. **Summary**  
   Given that both the validation and initial factor analysis yield the same number of factors and similar matrices it is without a doubt there are four factors that can be withdrawn from this dataset. Additionally, with this information a deep understanding of consumer behavior can be unravel and aid in the growth of a company.

# **APPENDIX I: OVERVIEW OF DATASET**

import pandas as pd

from pandas import Series, DataFrame

import numpy as np

import matplotlib as plt

import seaborn as sns

%matplotlib inline

df\_data = pd.read\_excel(r'HBAT(1).xls')

df = DataFrame(df\_data)

df = df.drop(columns = ['id', 'x1','x2','x3','x4','x5','x19','x20','x21','x22','x23' ])

print (df.columns)

df.columns =['x6 Product Quality', 'x7 E-commerce', 'x8 Technical Support', 'x9 Complaint Resolution','x10 Advertising',

'x11 Product Line', 'x12 Salesforce Image', 'x13 Competitive Pricing', 'x14 Warranty & Claimns', 'x15 Packaging',

'x16 Order & Billing', 'x17 Price Fexibility', 'x18 Delivery Speed']

df.head()

**APPENDIX II: Assessing the Appropriateness of Factor Analysis**

from pandas.plotting import scatter\_matrix

df\_matrix = pd.DataFrame(df)

corr = df\_matrix.corr()

corr.style.background\_gradient(cmap='RdYlGn')

**APPENDIX III: Deriving Factors and Assessing Overall Fit**

import pandas as pdÍ

from factor\_analyzer import FactorAnalyzer

import matplotlib.pyplot as plt

import sklearn.datasets

df\_matrix = df\_matrix.drop (columns = ['x15 Packaging', 'x17 Price Fexibility'])

fa = FactorAnalyzer()

fa.fit(df\_matrix,11)

eigen\_values, vectors = fa.get\_eigenvalues()

plt.title('Interpreting The Factors')

plt.xlabel('Factor')

plt.ylabel('Eigenvalue')

plt.scatter(range(1,df\_matrix.shape[1]+1),eigen\_values)

plt.plot(range(1,df\_matrix.shape[1]+1),eigen\_values)

plt.grid()

print("eigen\_values")

print (eigen\_values)

print("""

\*\*

""")

print("Variance")

print (fa.get\_factor\_variance())

print("""

\*\*

""")

**APPENDIX IV: Validation of Factor Analysis**

df\_matrix.shape

xdf = df\_matrix[:60]

xdf.shape

fa = FactorAnalyzer()

fa.fit(xdf,13)

eigen\_values, vectors = fa.get\_eigenvalues()

plt.title('Interpreting The NEW Factors')

plt.xlabel('Factor')

plt.ylabel('Eigenvalue')

plt.scatter(range(1,xdf.shape[1]+1),eigen\_values)

plt.plot(range(1,xdf.shape[1]+1),eigen\_values)

plt.grid()

eigen\_values