

# AUTONOMUS UNIVERSITY OF QUERETARO



## FACULTY OF ENGINEERING MASTER SCIENCE IN ARTIFICIAL INTELLIGENCE

### **Principal Component Analysis.**

Student:

Ing. Iván Alejandro García Amaya

Instructor:

Dr. Marco Antonio Aceves Fernández.

Course:

Machine Learning.

## Tabla de Contenido.

<b>1</b>	<b>Introduction.</b>	<b>2</b>
<b>2</b>	<b>Methodology</b>	<b>3</b>
2.1	Text. . . . .	3
2.2	Mean. . . . .	3
2.3	Center of data. . . . .	4
2.4	Covariance matrix. . . . .	4
2.5	Symmetric Power method. . . . .	4
2.5.1	Deflaction Techniques. . . . .	5
2.6	Principal Components. . . . .	5
2.7	min max normalization. . . . .	6
2.8	Z-Score Standardization. . . . .	6
2.9	Conditional mean imputation. . . . .	6
2.10	Regression imputation. . . . .	7
2.11	Prediction of new observation. . . . .	8
<b>3</b>	<b>Results and discussion.</b>	<b>9</b>
3.1	Imputation . . . . .	9
<b>4</b>	<b>Conclusion and Future Work</b>	<b>13</b>
<b>5</b>	<b>References.</b>	<b>14</b>

## Lista de Figuras.

1	No linear relationship (Douglas,2012). . . . .	9
2	linear relationship (Douglas,2012). . . . .	10
3	Dataset . . . . .	10
4	Subset. . . . .	11

## Lista de Tablas.

# **1 Introduction.**

In the following work an approach is made to a set of topics, such as Principal Components Analysis for dimension reduction in the text part, techniques of normalization such as max-min normalization and Z-Score Standardization, finally two imputation techniques, mean imputation which consists of replacing the missing values with their mean and the second technique Regression, in the dataset online\_retail\_II.

It is necessary to mention that this work has as a personal purpose, to understand the techniques in a deeper way, whether this purpose is fulfilled or not, will be a question that will be discussed later.

## 2 Methodology

### 2.1 Text.

For get a matrix W, it's necessary get two components, one, term frequency (tf) component, should depended upon the frequency with a query term occurs in a given document.

The second one, it's the inverse document frequency (idf), which measures the relative rarity of a term. It is usually given by. (Peter Jackson,2002).

$$idf_t = \log\left(\frac{N}{n_t}\right) \quad (1)$$

where N is the number of documents in the collection, and  $n_t$  is the number of documents in which term t appears.

The weight of a term, t, in a document vector, d, is given by

$$w_{t,d} = tf_{t,d} * idf_{t,d} \quad (2)$$

Where  $tf_{t,d}$  is a simple count of how many times t occurs in the document.

---

**Algorithm 1:** Text.

---

**input** : file.csv

**output:**  $W_{t,d}$

```
1 chart_vector  $\leftarrow$  Split_Clean(text);  
2  $tf_{t,d} \leftarrow$  term_frequency(vector_chart);  
3  $idf_{t,d} \leftarrow$  in_doc_freq(matriz_chart);  
4  $w_{t,d} \leftarrow$  get_weigths(tdt,d, idft,d);
```

---

### 2.2 Mean.

The mean of a random variable shows the location or the central tendency of the random variable.

$$\overline{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad (3)$$

### 2.3 Center of data.

It is necessary to extract the mean of each column of data, then subtract the mean from the same column, this is defined by the equation (4).

$$\hat{X} = X - \overline{X} \quad (4)$$

Where  $\hat{X}$  is the centralized data,  $X$  our data and  $\overline{X}$  is the mean of the data

### 2.4 Covariance matrix.

The covariance matrix is defined by the equation 5

$$C\hat{X} = \frac{1}{n} \hat{X}' \hat{X} \quad (5)$$

The  $ij^{th}$  element of  $C\hat{X}$  is the dot product between the vector of the  $i^{th}$  measurement type with the vector of the  $j^{th}$  measurement type. We can summarize several properties of  $C\hat{X}$ : (Jonathon Shlens, 2014)

- $C\hat{X}$  is a square symmetric  $m \times m$  matrix
- The diagonal terms of  $C\hat{X}$  are the variance of particular measurement types

### 2.5 Symmetric Power method.

A technique for the approximation of eigenvalues, is Power Method, defined like an iterative method, this is defined by the equation 6, to approximate an eigenvalue and an associated eigenvector of the  $n \times n$  matrix  $A$  given a nonzero vector  $x$ : (Burden R ,2005)

$$x_{n+1} = \frac{Ax_n}{||Ax_n||} \quad (6)$$

### 2.5.1 Deflation Techniques.

After having obtained the first eigenvalues, there is a set of techniques to obtain the others, one of these is deflation techniques. (Burden R ,2005)

This consists of creating a new matrix B and is defined by the equation 7

$$B = A - \frac{\lambda_1}{|v_1|^2} v_1 * v_1^T \quad (7)$$

Where  $v_1$  is our first eigenvectors,  $\lambda_1$  our first eigenvalue and A our data matrix

---

**Algorithm 2:** Power Method

---

**input** : Matriz A  
**output:** Eigenvectors  
**1 for**  $K \rightarrow n$  **do**  
**2**      $z_k = Aq_{k-1};$   
**3**      $q_k = z_k / |z_k|;$   
**4 end**

---

## 2.6 Principal Components.

Once the eigenvectors are obtained, it is necessary to multiply it by our data and it is defined by the equation 8.

$$PC = v' \hat{X} \quad (8)$$

Where v is a eigenvectors and X is a centralized data.

---

**Algorithm 3:** Principal Component Analysis.

---

**input** : Weights

**output:** Principal Components.

```
1 mean  $\leftarrow$  mean_d();  
2 covariance_matrix  $\leftarrow$  X * X';  
3 eigenvectors  $\leftarrow$  Power_method();  
4 % RFW : RowFeatureVector "traspose eigenvectors";  
5 % RDA : RowDataAdjust "mean-ajusted data transpose";  
6 Principal_Components  $\leftarrow$  RFW * RDA;
```

---

## 2.7 min max normalization.

Min - Max is a data normalization technique like Z score, decimal scaling, and normalization with standard deviation. It helps to normalize the data. It will scale the data between 0 and 1. This normalization helps us to understand the data easily.

$$X_n = (x - \min(x)) / (\max(x) - \min(x)) \quad (9)$$

## 2.8 Z-Score Standardization.

The standardize return a normalized value (z-score) based on the mean and standard deviation. A z-score, or standard score, is used for standardizing scores on the same scale by dividing a score's deviation by the standard deviation in a data set.

$$Z = (x - \mu) / \sigma \quad (10)$$

where  $\mu$  it's the mean and  $\sigma$  Standard Deviation

## 2.9 Conditional mean imputation.

Imputing Means Within Adjustment Cells. A common method in surveys is to classify nonrespondents and respondents into J adjustment classes, analogous to weighting classes, based on the observed variables, and then impute the respondent mean for nonrespondents in the same adjustment class. Assume equal probability sampling with constant sample weights, and let  $\bar{y}_{jR}$

be the respondent mean for a variable Y in class j. The resulting estimate of the mean of Y from the filled-in data is (Little R & Rubin D., 2020)

$$\bar{y}_{wc} = \frac{1}{n} \sum_{j=1}^J n_j \bar{y}_{jR} \quad (11)$$

where  $n = \sum_{j=1}^J n_j$  is the total sample size

## 2.10 Regression imputation.

The simple linear regression model, is a model with a single regressor x that has a relationship with a response y that is a straight line, we can found this equations in the book Introduction to linear regression Analysis (Douglas C. Montgomery, 2012).

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad (12)$$

where the intercept  $\beta_0$  and the slope  $\beta_1$  are unknown constants.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (13)$$

The mean of  $y_i$  is given by

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad (14)$$

and the mean of  $x_i$  by

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (15)$$

$\hat{\beta}_0$  is defined by the equation 16

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n y_i - \frac{(\sum_{i=1}^n y_i)(\sum_{i=1}^n x_i)}{n}}{\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}} \quad (16)$$

But, we can also define  $\hat{\beta}_0$  by equation 17

$$\hat{\beta}_0 = \frac{S_{xy}}{S_{xx}} \quad (17)$$

where  $S_{xy}$  is given by the equation 18



$$S_{xy} = \sum_{i=1}^n y_i(x_i - \bar{x}) \quad (18)$$

and  $S_{xx}$  by the equation 19

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 \quad (19)$$

## 2.11 Prediction of new observation.

An important application of the regression model is prediction of new observations  $y$  corresponding to a specified level of the regressor variable  $x$ . If  $x_0$  is the value of the regressor variable of interest, then (Douglas C. Montgomery, 2012).

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \quad (20)$$

is the point estimate of the new value of the response  $y_0$ .

$$\hat{y}_0 - t_{\alpha/2, n-2} \sqrt{MS_{Res} \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)} \leq y_0 \leq \hat{y}_0 + t_{\alpha/2, n-2} \sqrt{MS_{Res} \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)} \quad (21)$$

If we choose  $\alpha = 0.05$ , the critical value of  $t$  is  $t_{0.025, 18} = 2.101$ .

The quantity  $MS_{Res}$  is called the residual mean square. The square root of  $\hat{\sigma}^2$  is sometimes called the standard error of regression.

$$\hat{\sigma}^2 = \frac{SS_{Res}}{n-2} = MS_{Res} \quad (22)$$

The equation 23 show the error of squares  $SS_{Res}$ .

$$SS_{Res} = SS_T - \hat{\beta}_1 S_{xy} \quad (23)$$

$SS_T$  is the corrected sum of squares of the observations

$$SS_T = \sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n}; \quad (24)$$

regression sum of squares may be computed as

$$SS_R = \hat{\beta}_1 S_{xy}; \quad (25)$$

### 3 Results and discussion.

#### 3.1 Imputation

In order to be able to talk about the results obtained applying the model Linear Regression, it's necessary to mention testing Significance of Regression(Douglas,2012)

$$H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0 \quad (26)$$

These hypotheses relate to the significance of regression. Failing to reject  $H_0 : \beta_1 = 0$  implies that there is no linear relationship between x and y, As shown in the figure 1.

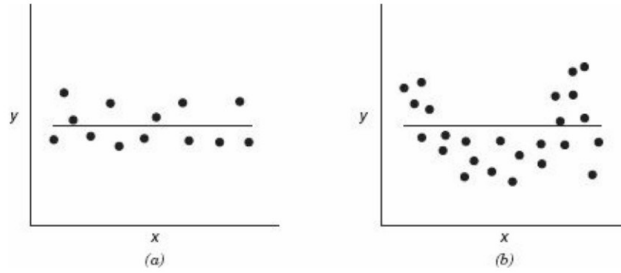


Figura 1: No linear relationship (Douglas,2012).

Alternatively, if  $H_0 : \beta_1 = 0$  is rejected, this implies that x is of value in explaining the variability in y. showed in figure 2

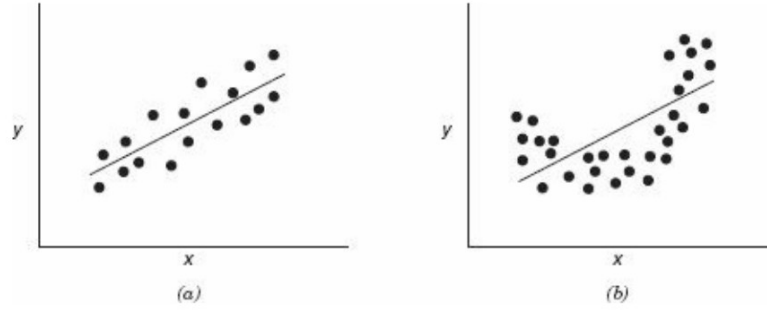


Figura 2: linear relationship (Douglas,2012).

As we can see in the figure 3, the dataset consists of 8 attributes and half a million characteristics, in addition to 2515 missing values, the first 100 values are within the 15000 characteristics.

Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer.ID	Country
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850	United Kingd
536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850	United Kingd
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850	United Kingd
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850	United Kingd
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850	United Kingd
536365	22752	SET 7 BABUSHKA NESTING BOXES	2	12/1/2010 8:26	7.65	17850	United Kingd
536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	12/1/2010 8:26	4.25	17850	United Kingd
536366	22633	HAND WARMER UNION JACK	6	12/1/2010 8:28	1.85	17850	United Kingd
536366	22632	HAND WARMER RED POLKA DOT	6	12/1/2010 8:28	1.85	17850	United Kingd
536368	22960	JAM MAKING SET WITH JARS	6	12/1/2010 8:34	4.25	13047	United Kingd
536368	22913	RED COAT RACK PARIS FASHION	3	12/1/2010 8:34	4.95	13047	United Kingd

Figura 3: Dataset

Example of missing value: 14341

Quantity : 3

Price : NA

Description : 'GREEN REGENCY TEACUP AND SAUCER'

Country : 'United Kingdom'

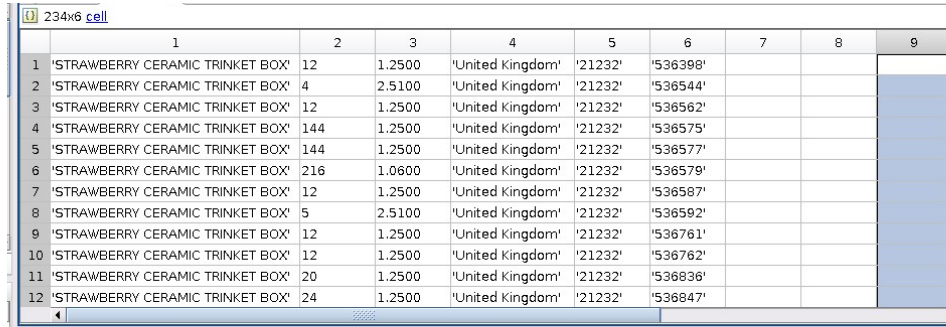
Stock Code : '15658'

As personal reasoning, an important step trying to find the linear relationship of the data, is to create subsets of data, the first depends on the

description of the data.

- We withdraw products with different description: 'SANDALWOOD FAN', 'POPCORN HOLDER'.
- We withdraw foreign countries such as: France, Germany, Iceland, Belgium, etc.
- Different Withdrawal Stock Code: '22197', '22190', '22212'.

After implementing this series of steps, the size of the subset decreases from 1014-770-700-234 as shown in figure 4.



	1	2	3	4	5	6	7	8	9
1	'STRAWBERRY CERAMIC TRINKET BOX'	12	1.2500	'United Kingdom'	'21232'	'536398'			
2	'STRAWBERRY CERAMIC TRINKET BOX'	4	2.5100	'United Kingdom'	'21232'	'536544'			
3	'STRAWBERRY CERAMIC TRINKET BOX'	12	1.2500	'United Kingdom'	'21232'	'536562'			
4	'STRAWBERRY CERAMIC TRINKET BOX'	144	1.2500	'United Kingdom'	'21232'	'536575'			
5	'STRAWBERRY CERAMIC TRINKET BOX'	144	1.2500	'United Kingdom'	'21232'	'536577'			
6	'STRAWBERRY CERAMIC TRINKET BOX'	216	1.0600	'United Kingdom'	'21232'	'536579'			
7	'STRAWBERRY CERAMIC TRINKET BOX'	12	1.2500	'United Kingdom'	'21232'	'536587'			
8	'STRAWBERRY CERAMIC TRINKET BOX'	5	2.5100	'United Kingdom'	'21232'	'536592'			
9	'STRAWBERRY CERAMIC TRINKET BOX'	12	1.2500	'United Kingdom'	'21232'	'536761'			
10	'STRAWBERRY CERAMIC TRINKET BOX'	12	1.2500	'United Kingdom'	'21232'	'536762'			
11	'STRAWBERRY CERAMIC TRINKET BOX'	20	1.2500	'United Kingdom'	'21232'	'536836'			
12	'STRAWBERRY CERAMIC TRINKET BOX'	24	1.2500	'United Kingdom'	'21232'	'536847'			

Figura 4: Subset.

If we look at the figure 4, it belongs to a subset of missing value data, this was achieved first through the description, all the different products in description were removed, in order to increase the linear relationship between the quantity and price data, in addition, we know that they are from the same country, this means that the price is marked for the owns currency, we do not have apparent problems of exchange currency, the third and last step is to eliminate the products with different Stock Code, then it assures us that the product has the same code, after applying this series of steps, what you might think is that the product in our subset are the same, only changed the quantity and the price, then we would think that it is relatively easy to calculate any value of Price depending on the Quantity.

The problem lies in the irregularity shown in the final data, as shown in the figure 4, we have the product, for example, in quantities of 144 pieces

at a price of 1.25 and on the other hand we observe that in quantities of 10 pieces with a price of 1.25, the same price. This apparent discrepancy in the data makes a linear relationship impossible, as mentioned at the beginning of the section.

---

**Algorithm 4:** Imputation.

---

```

input : dataset
output:  $y_0$ 

1  $r \leftarrow \text{get\_missing\_values}()$ ;
2  $len \leftarrow \text{length}(r)$ ;
3 for  $i \rightarrow len$  do
4    $Mv = r(i)$ ;
5    $country, sk, x_0 \leftarrow \text{get\_values}(Mv)$ ;
6    $subset \leftarrow \text{get\_set\_description}(Mv)$ ;
7    $subset_2 \leftarrow \text{clean\_country}(subset, country)$ ;
8    $subset_3 \leftarrow \text{clean\_stockCode}(subset_2, sk)$ ;
9    $y_0, a, b \leftarrow \text{imputation}(subset_3, x_0)$ ;
10     $y_0 \leftarrow \text{Mean}(subset_3, x_0)$ ;
11     $a, b \leftarrow \text{Linear\_Regression}(subset_3, x_0)$ ;
12     $a, b \leftarrow \text{Prediction}(y, \bar{x}, n, Sxy, Sxx, \hat{\beta}_0, \hat{\beta}_1, x_0)$ ;
13 end

```

---

## 4 Conclusion and Future Work

Within the first 100 subsets generated by from the missing values, it didn't found a linear relationship, therefore, is necessary to review new techniques, as we mentioned in the previous section, techniques such as mean imputation and mode imputation but you can explore techniques like Logistic Regression, this is more of a personal opinion, but it will be necessary to further explore the entire Regression technique, seeking greater rigor in the investigation of the Regression area.

For this work, in order to better understand the nature of PCA, the Power Method was implemented, the results obtained are not entirely exact, it is necessary to further investigate the nature of eigenvectors and delve deeper into the area of linear algebra. This means, in order to be more rigorous in the results obtained, it is necessary to delve deeper into this area.

## 5 References.

Douglas C. Montgomery. (2012). Introduction to Linear Regression Analysis. Hoboken, New Jersey.: John Wiley & Sons, Inc.

Jonathon Shlens. (2014). A Tutorial on Principal Component Analysis. Google Research.

Lindsay I Smith. (2002). A tutorial on Principal Components Analysis. Otago, New Zealand: University of Otago.

Peter Jackson . (2002). Natural Language Processing for Online Application. Wolverhampton WV1 1SB, United Kingdom: University of Wolverhampton.

Richard L. Burden & J. Douglas Faires. (2005). Numerical Analysis. United States of America: Thomson Higher Education.

Roderick J. A. Little & Donald B. Rubin. (2020). Statical Analysis with Missing Data. USA: John Wiley & Sons, Inc.