A Methodology for Classifying Visitors to an Amusement Park

VAST Challenge 2015 - Mini-Challenge 1

Gustavo Dejean *, *Departamento de Computación*

Universidad Nacional del Oeste - San Antonio de Padua, Pcia. de Buenos Aires – Argentina.

ABSTRACT

The main contribution of this work is showing how to obtain a classification of visitors to an amusement park by using cluster analysis and visualization techniques. The selection of variables for K-means algorithm and the results obtained are visually analyzed in dispersion graphs according to their Principal Components, in boxplots and in a Linear Model so as to fine-tune a result that can explain differences between the groups and similarities within each group.

Keywords: *visual analysis, vast challenge, data mining, clustering, K-means, cluster analysis.*

1 INTRODUCTION

Frequently, clustering algorithms identify clusters that do not exist in reality [1] or trivial clusters that are not relevant for researcher's goals. That is why, it is important to be able to visualize results. This is attained both with 2D and 3D diagrams with their first Principal Components (PC), as well as with other types of diagrams.

2 THEORY

2.1 Material

Three archives of moves were used, plus a fourth one that completed Friday's missing records. Overall, they contain 26,021,945 records. Each record accounts for a move or a check-in from a visitor at the park on a specific day and time. Archives are included in [2]. Tools that were used include: Postgresql 8.4 database, IBM SPSS Statics v22; JMP 11 (SW), and Tableau 8.2.

2.2 Methods

The process used for reaching to a final cluster may be divided into four stages, that were executed concurrently with different intensities, according to the progress of the work. These four stages were: 1) Creation of new variables; 2) Creation of models and visualization; 3) Principal Components (PC); 4) K-means and cluster visualization.

2.2.1 Creation of New Variables

Approximately 64 new variables were created. Some of them were instrumental for developing a Linear Model that enabled visualizing and understanding the data and outcomes that were obtained. Table 1 describes the main variables that were created.

2.2.2 Creation of Models and Visualization

The Linear Model (LM) exhibited in Figure 1 was instrumental for visualizing the outcome of the final clustering.

Table 1: List of the main variables or groups of variables created for each id

variable name	clarification	Input de K-means
count_of_movement	totals in the three day	no
count_of_check-in-entry		no
count_of_check-in-play		no
time_of_permanencia		no
time_average		no
count_for_each_items	of check-in	yes
count_for_each_rubro		no
max, min and avg for each rubro		some

The LM shows a linear correlation between the number of moves vs. the amount of check-ins in the games (for every ID), yielding R^2 = 0,848. An interesting observation of the LM is that, in general, each point represents a group of visitors. That is, each group of persons entering the park through the same door and at the same time was represented in the LM by points that were overlapping or very close to one another. It will always be interesting visualizing any possible outcome from a clustering process by using this Model.

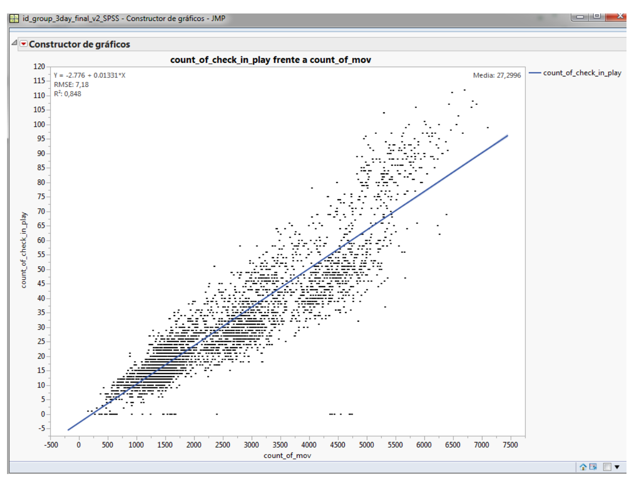


Figure 1: Linear Model. count_of_ movement vs. count_of_check-in.

* email: dejean2010@gmail.com

By observing the outlier points in the LM, 39 IDs with zero check-in in the games were identified. These IDs were grouped in Cluster 0; and they were not taken into account when applying the K-means algorithm nor in any other LM tunings, nor in the analysis of PC. In Cluster 0, this would be represented mainly by the park employees. The LM without outliers remained with a R^2 =0,854.

2.2.3 *Principal Components*

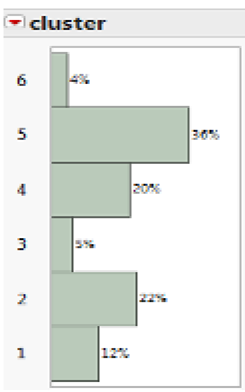
In the case of the study, as the dimensionality of the problem was greater than three, the PC were used for reducing it and, thus, for visualizing the outcomes in 2D or 3D graphs, by using the first two or three PC.

2.2.4 *K-means and cluster visualization*

An important point in this stage is determining which is the appropriate amount for visitor classification. The CCC criterion was used: Sarle's Cubic Clustering Criterion [3] taking the first peak with CCC >3. The other important point is selecting the variables for the input of the K-means algorithm that may yield a better grouping.

3 *RESULTS*

With the described procedure, for a subset of variables, a K = 6 was obtained. Thus, in the final result, 7 clusters are obtained. (N.B. Bear in mind that cluster 0 had been created), see figure 2.



cluster	Count id	Count group (approximate)	% id
6	490	176	4,31
5	4038	909	35,51
4	2305	906	20,27
3	606	174	5,33
2	2488	347	21,88
1	1407	371	12,37
0	39	3	0,34
total	11334	2886	100%

Figure 2: Distribution of each Cluster. Cluster 0 is not shown in the histogram.

The interpretation of each group was done by using boxplots, as shown in Figures 3. For example, when comparing the cluster, it was discovered that cluster 1 and cluster 3 show preferences for Thill Rides games and rejection for Kidder Rides. Similarly, with the other variables, in general, relevant differences were obtained between the clusters, thus accounting for the classification that was obtained.

Another interesting visualization of the results is shown in Figure 4, where the LM is used for visualizing each cluster. It should be noted, for example, that it is easy to read and detect which clusters have compulsive players and which have very low profile players, among other features. Overall, each cluster has a different straight line tangent. The number of groups in each cluster was roughly obtained with a SQL query grouped per time of entry and per entry point.

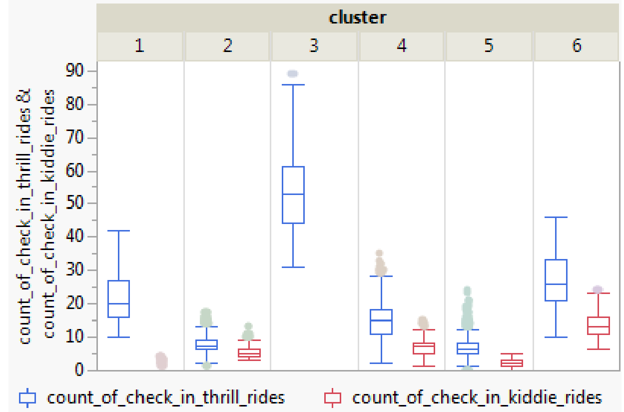


Figure 3: Preferences for Thill Rides and rejection for Kidder Rides for cluster 1 and 3

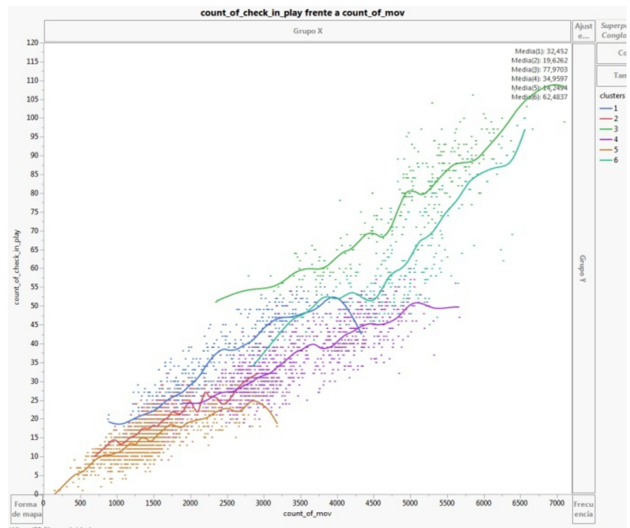


Figure 4: Visualization of Clusters in the LM Rides

4 *CONCLUSION*

In the resulting classification, it was possible to verify the existence of similarities between the visitors of each cluster and the differences between different clusters, mainly using boxplots, but also leveraging interpretation with the LM visualization.

ACKNOWLEDGMENTS

To the authorities of Universidad Nacional del Oeste and specially to Mg. Antonio Fotti and Lic. Ariel Aizemberg for enabling the research task.

REFERENCES

[1] Johnson Dallas E., Métodos Multivariados Aplicados al análisis de Datos, page 321. México, International Thomson Editores, 2000.

[2] http://www.vacommunity.org/2015+VAST+Challenge %3A+MC1

[3] Sarle, W.S., The CuvicClustering Criterion, SAS Technical Report A-108, Cary NC: SAS Institute, Inc., 1983.