

## Article

# A Hospital Recommendation System Based on Patient Satisfaction Survey

Mohammad Reza Khoie <sup>1,\*†</sup>, Tannaz Sattari Tabrizi <sup>1,\*†</sup>, Elham Sahebkar Khorasani <sup>2</sup>,  
Shahram Rahimi <sup>1</sup> and Nina Marhamati <sup>1</sup>

<sup>1</sup> Department of Computer Science, Southern Illinois University, Carbondale, IL 62901, USA; rahimi@cs.siu.edu (S.R.); nina@siu.edu (N.M.)

<sup>2</sup> Department of Computer Science, University of Illinois at Springfield, Springfield, IL 62703, USA; esahe2@uis.edu

\* Correspondence: rezakhoie@siu.edu (M.R.K.); tannaz@siu.edu (T.S.T.); Tel.: +1-618-303-1088 (M.R.K.); +1-618-303-0536 (T.S.T.)

† These authors contributed equally to this work.

Received: 28 July 2017; Accepted: 11 September 2017; Published: 21 September 2017

**Abstract:** Surveys are used by hospitals to evaluate patient satisfaction and to improve general hospital operations. Collected satisfaction data is usually represented to the hospital administration by using statistical charts and graphs. Although such visualization is helpful, typically no deeper data analysis is performed to identify important factors which contribute to patient satisfaction. This work presents an unsupervised data-driven methodology for analyzing patient satisfaction survey data. The goal of the proposed exploratory data analysis is to identify patient communities with similar satisfaction levels and the major factors, which contribute to their satisfaction. This type of data analysis will help hospitals to pinpoint the prevalence of certain satisfaction factors in specific patient communities or clusters of individuals and to implement more proactive measures to improve patient experience and care. To this end, two layers of data analysis is performed. In the first layer, patients are clustered based on their responses to the survey questions. Each cluster is then labeled according to its salient features. In the second layer, the clusters of first layer are divided into sub-clusters based on patient demographic data. Associations are derived between the salient features of each cluster and its sub-clusters. Such associations are ranked and validated by using standard statistical tests. The associations derived by this methodology are turned into comments and recommendations for healthcare providers and patients. Having applied this method on patient and survey data of a hospital resulted in 19 recommendations where 10 of them were statistically significant with chi-square test's *p*-value less than 0.5 and an odds ratio z-test's *p*-value of more than 2 or less than −2. These associations not only are statistically significant but seems rational too.

**Keywords:** health data analytics; survey analysis; HCAHPS; hospital consumer assessment of healthcare providers and systems; unsupervised learning

## 1. Introduction

Patient satisfaction has been proven to be one of the most valid indicators of the quality of care. Analysis of patient satisfaction data is in demand by many health-care providers. Most health-care providers, from doctor's offices to clinics and hospitals, collect patient satisfaction surveys to evaluate their various services and patient experience. This increasingly growing data is conventionally analyzed by statistical methods, such as analysis of variance (ANOVA) [1], simple regression, Fisher's approach and extensions [2], Neyman's approach to randomization-based inference [2], etc. Such methods typically approach the problem with a specific question in mind and find the relation between one or more independent variables and a dependent variable. For example, they compute the

percentage of the patients that have rated each hospital's services similarly or at most provide some correlations between specific groups of patients and their answers to a specific satisfaction question.

For improving patient satisfaction, issues of health care provided at the hospital level and the factors that originate those issues from patients' point of view should be discovered. Therefore, survey data should be either manually analyzed by examining each possible pattern in the data set using conventional methods or an unsupervised methodology is needed to do the analysis with least amount of human interaction. Such methodology should get the satisfaction survey data, find patterns that are repeated among patients' demographics and their satisfaction level in different fields, validate the patterns and compile them into a set of recommendations to help hospitals improve satisfaction within various patient communities.

To this end, a new hybrid methodology is proposed that differs from these conventional approaches in that it is not bound to a single outcome or dependent variable. The focus of this approach is to find patterns in patients' responses to all satisfaction questions and relate them to patients' demographics. The proposed methodology is focused on discovering issues of the health care provided at the hospital level and the factors that originate those issues from the patients' point of view. This methodology is a hybrid unsupervised clustering-labeling method, which finds associations between various levels of patients' satisfaction and demographics. The associations are validated by using standard statistical models and turned into useful recommendations for hospitals in order to improve patients' experience, save cost, and build long-term patient loyalty. The methodology can be generalized to any complex multi-level survey analysis.

The article is organized into eight sections. The next section describes the standard survey instrument used for collecting patient satisfaction data. Section 3 reviews the literature on the analysis of hospital survey data as well as modern survey data analysis methods. The proposed methodology for the analysis of the survey data will be presented in Section 4. Section 5 reports on the experimental results of the analysis using the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) dataset. The validation of the results of the analysis are discussed in Section 6. Section 7 explains how to convert the associations derived from the analysis into recommendations for healthcare providers. The last section draws a conclusion and future direction of this study.

## 2. Literature Review

This section describes the HCAHPS dataset and briefly reviews modern survey data analysis methods and their shortcomings in analyzing the HCAHPS dataset.

### 2.1. HCAHPS Hospital Survey Data

HCAHPS [3] is a standard survey instrument used by many hospitals to evaluate patients' experience. This data is provided by the HCAHPS database, which is funded by U.S. agency for health care research. The centers for Medicaid and Medicare services use the scores from HCAHPS to reimburse hospitals for patient care. Providing a high quality care is directly related to a hospital's revenue and many hospitals are looking for ways to improve patient experience and achieve a higher HCAHPS score.

Table 1 gives a brief description of the satisfaction questions on the HCAHPS survey instrument and the categories that they fall into. As shown in the table, the survey questions are divided into six sections where each section has a number of multiple choice questions. The number of choices for each question is also specified in Table 1. For instance, the section on "care from doctor" measures patient satisfaction with the care provided by doctor(s) using three questions about doctor's respect, listening, and explaining. Each question has four choices (Never, Sometimes, Usually, and Always).

**Table 1.** Satisfaction Questions Categories.

<i>Section</i>	<i>Number of Question in Each Section</i>	<i>Number of Choices for Each Question</i>
<i>Care from Nurses</i>	5	4
<i>Care from Doctors</i>	3	4
<i>Hospital Environment</i>	2	4
<i>Experience in Hospital</i>	5	4
<i>When You Left</i>	1	2
<i>Overall Rating</i>	1	2
	1	11

Table 2 presents the types of demographic questions in HCAHPS survey instrument. All demographic identifiers (except for “age” and “discharge date”) are categorical. The HCAHPS survey questionnaire is brought in the Appendix A.

**Table 2.** Demographic Data Types.

<i>Questions</i>	<i>Number of Choices</i>
<i>Age</i>	-
<i>Discharge Date</i>	-
<i>State</i>	44
<i>Racial Category</i>	6
<i>Overall Health</i>	5
<i>Education</i>	6
<i>Ethnicity</i>	5
<i>Patient-Filled Race</i>	5
<i>Patient’s Language</i>	3
<i>Gender</i>	2
<i>Principal Reason</i>	3
<i>Admission Source</i>	10
<i>Survey Language</i>	2

## 2.2. Review of Existing Studies on HCAHPS Dataset

There have been several studies on the HCAHPS dataset. Stratford [4] defined a number of objectives to extract useful knowledge from the HCAHPS survey data and studied the effect of such knowledge on hospital care improvement.

Sheetz et al. [5] investigated the relationship between postoperative morbidity and mortality and patients’ perspectives of care in surgical patients. In their article, the overall satisfaction score is used along with Michigan Surgical Quality Collaborative clinical registry as a measure of patients’ perspective of care.

Quite a few studies have explored specific relationships between a single satisfaction question and one or more of patients’ demographic information. Goldstein et al. [6] conducted an analysis of racial/ethnicity in patients’ perceptions of inpatient care. Using regression, they concluded that non-Hispanic Whites on average tend to go to hospitals that deliver better patient experiences to all patients as compared to the hospitals that are typically used by African American, Hispanic, Asian/Pacific Islander, or multiracial patients [6].

Elliot et al. analyzed the association of gender with different aspects of satisfaction, [7] and, in a separate study, analyzed hospital ranking variation with patient health status and race/language and slightly with patient’s education and age [8].

Klinkenberg [9] explored the relation between the willingness to recommend the hospital and other satisfaction identifiers. This paper discovers that hospitals that focus resources on improving

interpersonal aspects of care such as nurses and doctors' courtesy, respect, listening, room cleanliness, etc. will be most likely to see improvements in satisfaction scores. The paper does not consider patients' demographic data.

The existing literature on analysis of the HCAHPS dataset is mostly hypothesis-driven and only considers specific aspects of patient satisfaction or demographics. In contrast, the methodology presented in this paper does not assume any specific hypothesis. Instead, we run a data-driven exploratory analysis which inspects all aspects of patient satisfaction as well as patient demographics and discovers interesting associations in the HCAHPS dataset.

### *2.3. Shortcomings of Existing Survey Analysis Methods*

In addition to the literature on analysis of the HCAHPS dataset; it is also worth reviewing the methods typically used for general survey analysis. Commonly used exploratory data analysis methods such as ANOVA, regression, discriminant analysis, and factor analysis are not applicable to HCAHPS data because of its unique characteristics.

Analysis of variance (ANOVA) is a collection of statistical methods that form an exploratory tool for explaining observations. ANOVA provides a statistical test of whether or not the means of several groups are equal [2]. For finding a specific correlation in the HCAHPS dataset, different levels of ANOVA should be combined. To this end, all possible combinations of satisfaction questions and demographic data should be exhaustively tested, which could be time prohibitive. In addition, ANOVA assumes a normal distribution of the sample observation, continuous dependent variables, and at least one categorical independent variable with two or more levels. The sample data collected for HCAHPS is not guaranteed to be normally distributed. Moreover, the demographic data which form the dependent variables are not always categorical. Given the violation of these assumptions, ANOVA may not produce reliable results for HCAHPS dataset.

Regression analysis is a statistical tool for investigation of relationship between multiple continuous or categorical independent variables and a continuous dependent variable [10]. Regression models can be used for validating correlations, although they are usually used for predicting and forecasting. Variations of this model can be used for the HCAHPS categorical dependent variables and lead to nonlinear models. Using these interpretation, different hypothesis on the data set can be tested by using forward and backward selection of combinations of satisfaction questions and patients' demographic data. However, forward and backward selection is not efficient in the context of HCAHPS data as the search space contains a very large number of combinations of dependent and independent variables. Moreover, fitting a separate model for each satisfaction identifier cannot capture the relationships between different satisfaction questions.

Discriminant analysis is another technique that allows for studying the difference between two or more groups of objects with respect to several variables simultaneously [11]. This model can be fitted to HCAHPS data set in terms of categorical dependent variables as the satisfaction questions and continuous and categorical independent variables such as patient's demographical data. Although this method works better than regression in terms of interpreting categorical variables, it suffers from the same deficiencies when it comes to analyzing the HCAHPS dataset.

Factor analysis is another method that is typically applied to survey data. The main application of this method is to reduce the number of variables and to detect structure in the relationship between variables. In particular, factor analysis can be used to explore the data for patterns, confirm hypotheses, or reduce a large number of variables to a more manageable number [12]. Compared to regression and discriminant analysis, factor analysis is more suitable for an exploratory analysis of the HCAHPS dataset as it does not require a priori hypothesis; however, it has two limitations: (1) the naming of the factors can be problematic and may not accurately reflect the variables within that factor. In particular, it may not be possible to directly compile factors into a set of recommendations for the hospitals. (2) factor analysis is based on the assumption that there is a linear relationship between factors and the

variables when computing correlations. The features in the HCAHPS survey data may not necessarily be linearly correlated and factor analysis cannot capture non-linear relations.

### 3. Analyzing HCAHPS Data

This section describes the proposed methodology for analyzing the HCAHPS data. The analysis is done in three steps: 1—data preparation, 2—two-layer cluster analysis, and 3—salient feature extraction and associations.

#### 3.1. Data Preparation

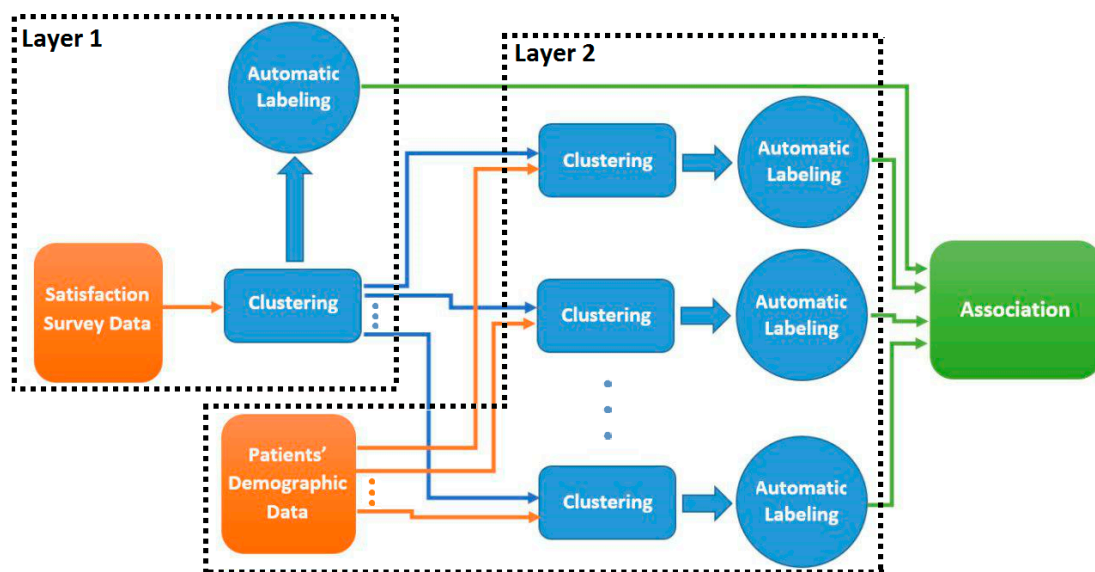
The first problem that should be handled in data preparation is the nature of some questions, called skip questions. A skip question by itself does not provide any information about a patient but it determines whether some other questions, called dependent questions, are applicable to the patient. For instance, a skip question inquires if the patient has used the bathroom or not. If not, the patient skips all of the dependent questions related to the bathroom cleanliness. Since skip questions, by themselves, do not provide any data about a patient, it would be reasonable to omit them from the dataset and treat their empty dependent questions as missing values. For example, if a patient has not used the bathroom, all of the bathroom-related questions have missing value for that patient.

There are two basic approaches for handling missing values: (1) the complete case analysis which ignores the records with missing data, or (2) the imputation of missing values. The imputation method can be further divided into single imputation where each missing value is replaced by a single value, and multiple imputations where each missing value is replaced by multiple values to reflect the uncertainty. The complete case analysis could introduce a selection bias and may lead to loss of information. The added bias makes the data related to those questions more homogenous. If all of the patients have similar opinion on the related questions the added bias will make those questions insignificant. However, if in fact there are two opposite opinions on this matter that are separated based on the demographic features of patients, the separation will be more statistically significant. Therefore, a single imputation method is applied in this study for handling missing values in HCAHPS dataset. This approach may reduce the variance and add bias to range of the imputed variables but will not result in loss of information. The reduction in the variance of an imputed variable may decrease the chance of that variable being selected as a significant feature when compared to other variables, but there is still a chance of being selected if the data is extensively divided on that variable.

For imputing the missing values, the *K*-nearest neighbor imputation method (KNNI) [13] is used. This method is chosen because it is applicable to both continuous and categorical variables and it can be applied to the data set automatically and without any supervision. This method imputes the missing value based on the *K* nearest neighbors of the record with missing value. For categorical features, the missing value is replaced with the category which has the majority in the *K* nearest neighbors. For continuous features, the missing value is replaced with the weighted average of that feature in the nearest neighbors. To have minimum computational complexity,  $K = 3$  has been chosen for KNNI as it is the least number of neighbors that could produce reasonable results for finding majority of classes among categorical variables.

#### 3.2. Two-Layer Cluster Analysis

The goal of this step is to do exploratory data analysis in order to find hidden patterns in the HCAHPS dataset and to identify the main sources of patient dissatisfaction. To this end, two layers of clustering is performed on HCAHP data as illustrated in Figure 1.



**Figure 1.** Two-layered analysis method. In the first layer, the clusters and their labels are generated based on the satisfaction questions. In second layer, the clusters of first layer are re-clustered based on patient demographic data.

In the first layer, patient data is clustered based on satisfaction questions (listed in Table 2) to group the patients with similar satisfaction identifiers. We examined various clustering methods such as K-means, DBScan, and Spectral clustering and ultimately decided to choose K-means for its simplicity and efficiency. K-means is the most commonly used clustering method with two main problems. First, it is sensitive to the initialization of the cluster centers and might converge into a local optimum. Second, it requires a pre-specified number of clusters ( $k$ ). To address the first problem, K-Means++ is typically used to initialize the cluster centroids before proceeding with the standard K-means. With K-means++ initialization, the algorithm is guaranteed to find optimal clusters with  $O(\log k)$  competitive with K-means optimal solution [14]. For solving the second problem, there are several techniques offered to extract the optimal number of clusters. The Calinski-Harabasz criterion method is chosen for the estimation of number of clusters for K-means. This method finds the best number of clusters by applying the criterion of minimum within cluster sum of squares. This procedure ensures an effective reduction of the number of possible splits [15], which prevents overfitting in association extraction procedure. This method is implemented in R using the *vegan* v2.4-2 package by Jari Oksanen based on the Algorithm 1.

---

**Algorithm 1.** Optimal number of clusters using Calinski-Harabasz criterion.

---

```

Find Number of Clusters (data, minNumClusters = 1, maxNumClusters = 10 )
Fit ← cascadeKM (data, inf.gr = minNumClusters, sup.gr = maxNumClusters, iter = 100, criterion = "calinski")
calinski_best ← which.max (fit.results [2,])
Return calinski_best
  
```

---

To apply Kmeans++ algorithm on the HCAHPS data with mixed categorical and continuous variables, the categorical variables are transformed to dummy variables and the continuous variables are normalized using z-score normalization. By this transformation all the variables are in one spatial distance range, which is suitable for applying the Kmeans++ algorithm.

After applying Kmeans++ with selected number of clusters, the salient features of each cluster are derived using automatic cluster labeling to mark the important features that make up a cluster.



The salient features of a cluster are the ones whose values are significantly different (in a statistical sense) in the cluster compared to those in the other clusters.

The clusters of layer one are then fed into the second layer for further analysis. Each satisfaction cluster of the first layer is clustered again; but this time based on the demographic features of each record listed in Table 2 (e.g., patient's age, race, etc.). The salient features of each sub-cluster are then derived to find the important features that make up a sub-cluster.

### 3.3. Salient Feature Extraction

One can draw associations between the salient features of the outer (satisfaction) cluster and the salient features of its inner (demographic) sub-clusters. For instance, suppose that as a result of the first layer we get an outer cluster whose salient features indicate low values for "satisfaction with Doctor". This cluster is then further clustered into demographic sub-groups. Suppose that the salient features of one of the sub-groups indicates higher values for age and a particular doctor (Doctor X) who visited most patients in this sub-group.

Putting the salient features of a cluster and its sub-clusters together, one can draw an association between older patients who expressed low satisfaction with their doctor and who were visited by Doctor X. Such associations must be further validated through statistical evaluations and can be used to make recommendations to the hospital. For example, the recommendation system might recommend not to assign Doctor X to older patients.

We extract the salient features of each cluster based on the methodology proposed in [16]:

1. The centroid of a cluster  $k$  is computed as the average of the points in the cluster:

$$X_k = \frac{\sum_{i=1}^N P_i}{N} \quad (1)$$

where  $X_k$  is the centroid of cluster  $k$  and  $P_i$  is a point in cluster  $k$ .

2. The Euclidean distance of each point to its cluster centroid is computed:

$$d_i = \sqrt{\sum_j (P_{ij} - X_{kj})^2} \quad (2)$$

3. The points in each cluster are divided into *in-pattern* and *out-pattern* records. The records whose distance lie within the range defined by (3) are called in-pattern records while all other records including the ones in other clusters are called out-pattern records.

$$\mu_k - z\sigma_k < d_i < \mu_k + z\sigma_k \quad (3)$$

where  $\mu_k$  and  $\sigma_k$  are the mean and standard deviation of the points in cluster  $k$ , respectively, and  $z$  is a constant factor. Smaller  $z$  results in more out-pattern records and larger  $z$  result in more in-pattern records.

4. For each feature  $v$  and cluster  $k$ , the mean of all in-pattern records,  $\mu_{in}(k, v)$  and the mean of the out-pattern records,  $\mu_{out}(k, v)$ , are computed:

$$\mu_{in}(k, v) = \frac{\sum_{P_i \in \varphi_{in}} P_{iv}}{|\varphi_{in}(k)|} \quad (4)$$

$$\mu_{out}(k, v) = \frac{\sum_{P_i \in \varphi_{out}} P_{iv}}{|\varphi_{out}(k)|} \quad (5)$$

where  $\varphi_{in}(k)$  and  $\varphi_{out}(k)$  are the set of in-pattern and out-pattern points in cluster  $k$ , respectively.

5. A difference factor,  $df(k, v)$ , is calculated for each feature  $v$  in cluster  $k$  based on Equation (6):

$$df(k, v) = \frac{\mu_{in}(k, v) - \mu_{out}(k, v)}{\mu_{out}(k, v)} \quad (6)$$

6. The mean and standard deviation of the difference factors for all features in cluster  $k$  are calculated as follows:

$$\mu_{df}(k) = \frac{\sum_{v=1}^D df(k, v)}{D} \quad (7)$$

$$\sigma_{df}(k) = \sqrt{\sum_{v=1}^D (df(k, v) - \mu_{df}(k))^2 / D} \quad (8)$$

where  $D$  is the number of features in the input space.

7. A feature  $v$  is a salient feature in cluster  $k$  if its corresponding difference factor in  $k$  deviates considerably from  $\mu_{df}(k)$ . More formally, feature  $v$  is a salient feature in cluster  $k$  if:

$$df(k, v) \leq \mu_{df}(k) - z\sigma_{df}(k) \quad (9)$$

$$df(k, v) \geq \mu_{df}(k) + z\sigma_{df}(k) \quad (10)$$

where  $z$  is a constant factor. The smaller the  $z$  the more salient features in each cluster. Salient feature extraction method is outlined in Algorithm 2.

---

**Algorithm 2. Salient features extraction.**

---

*FindingSalientFeatures* ( $noc$ ,  $clustered\_data$ ,  $z$ )

*Comment:* calculating the center of each clustering by averaging records in the cluster.

For  $i \leftarrow 0$  to  $noc - 1$

$cluster\_centers[i] \leftarrow \text{average over columns}(clustered\_data[i])$

*Comment:* calculating distance of records from their assigned cluster center.

For  $i \leftarrow 0$  to  $noc - 1$

    For  $j \leftarrow 0$  to  $\text{length}(clustered\_data[i]) - 1$

$distance\_matrix[i][j] \leftarrow \text{distance}(clustered\_data[i][j], cluster\_centers[i])$

*Comment:* calculating the average distance of each cluster from its center.

For  $i \leftarrow 0$  to  $noc - 1$

$average\_distance[i] \leftarrow \text{average over columns}(distance\_matrix[i])$

*Comment:* calculating standard deviation of distances in each cluster.

For  $i \leftarrow 0$  to  $noc - 1$

$standard\_deviation[i] \leftarrow \sqrt{\text{average over } j((distance\_matrix[i][j] - Average\_distance[i])^2)}$

*Comment:* Finding in pattern and out pattern records in each cluster

$counter \leftarrow 0$

For  $i \leftarrow 0$  to  $noc - 1$

    For  $j \leftarrow 0$  to  $\text{length}(clustered\_data[i]) - 1$

        If  $(distance\_matrix[i][j] < (average\_distance[i] + (z * standard\_deviation[i])))$

        AND  $distance\_matrix[i][j] > (average\_distance[i] - (z * standard\_deviation[i]))$

$In\_patterns[i][counter] \leftarrow j$

$counter++$

*Comment:* Calculating the mean of each feature in each cluster for in-pattern neurons.

For  $i \leftarrow 0$  to  $noc-1$

    For  $j \leftarrow 0$  to  $\text{length}(clustered\_data[i]) - 1$

        If  $(j \text{ in } In\_patterns[i])$

$In\_pattern\_mean[i] \leftarrow In\_pattern\_mean[i] + clustered\_data[i]$

        else



**Algorithm 2. Cont.**


---

```

        Out_pattern_mean[i] ← Out_pattern_mean[i] + clustered_data[i]
    In_pattern_mean[i] ← In_pattern_mean[i]/length(In_patterns[i])
    Out_pattern_mean[i] ← Out_pattern_mean[i]/length(Out_patterns[i])
    Comment: Calculating the difference factor of in and out pattern records.
    For i ← 0 to noc − 1
        For j ← 0 to number_of_columns(clustered_data) − 1
            difference_factor[i][j] ← In_pattern_mean[i][j] − Out_pattern_mean[i][j]
    Comment: Calculating the mean difference factor of each dimension.
    Mean_difference_factor ← average over row(difference_factor)
    Comment: Calculating the standard deviation difference factor of each cluster.
    For i ← 0 to noc − 1
        For j ← 1 to number_of_columns(clustered_data) − 1
            Difference_factor_SD[i] ← Difference_factor_SD[i] + difference_factor[i][j] −
            Mean_difference_factor[i]^2)
            Difference_factor_SD[i] ← sqrt(Difference_factor_SD[i]/
            number_of_columns(clustered_data))
    Comment: Calculating a matrix of salient dimentions.
    For i ← 0 to noc − 1
        For j ← number_of_columns(clustered_data) − 1
            If (difference_factor[i][j] ≤
            (Mean_difference_factor[i] − (z* Difference_factor_SD[i])))
                Salient_dimension[i][j] = −1
            else if (difference_factor[i][j] ≥
            (Mean_difference_factor[i] + (z* Difference_factor_SD[i])))
                Salient_dimension[i][j] = 1
    Return Salient_dimension

```

---

To illustrate the extraction of salient features, suppose, as an example, that we have a dataset with five features as shown in Table 3.

**Table 3.** Features of Satisfaction Dataset.

Features	Description
D1	Communication with doctor
D2	Communication with Nurse
D3	Pain Management
D4	Cleanliness
D5	Quietness

Suppose that the data points of this feature space are clustered into three groups with the centroids listed in Table 4.

**Table 4.** Clusters Centroids.

Features	C1	C2	C3
D1	−0.0126243	0.97095961	−0.9913867
D2	0.03791055	−0.07749729	0.02231231
D3	−1.1444478	0.8585494	0.8681821
D4	−1.1509251	0.8681547	0.8681547
D5	−0.00279281	0.009980126	−0.0060900

To find the in-pattern and out-pattern records in the cluster, the distances of each record to all three cluster centroids are computed. In addition, the mean and standard deviation of distances for

each cluster centroid are calculated. If a record's distance from a cluster centroid is within one standard deviation from the mean, then it is considered an in-pattern record of that cluster. Otherwise, it is an out-pattern record of the cluster.

Suppose that there are nine records in the dataset. Table 5 shows the distance of each record to all three cluster centroids as well as the mean and standard deviation of each cluster. The in-pattern records of each cluster are shown in bold.

**Table 5.** Distance of records from centroid. Bold numbers denote in-pattern records which are in range of one standard deviation from the mean.

Records	C1	C2	C3
R1	<b>1.967133</b>	0.7958750	<b>1.3957300</b>
R2	2.121455	<b>1.6483554</b>	0.9116866
R3	<b>1.675209</b>	<b>1.5959835</b>	<b>1.7050408</b>
R4	<b>1.518996</b>	<b>1.6174912</b>	<b>1.5210744</b>
R5	<b>1.697462</b>	<b>1.0558556</b>	2.1432977
R6	<b>1.843458</b>	0.8124378	<b>1.1173187</b>
R7	1.191784	2.2571117	<b>1.3717867</b>
R8	2.266916	0.7501202	<b>1.4390353</b>
R9	<b>1.706186</b>	<b>1.6242395</b>	<b>1.7090141</b>
R10	1.107229	2.0101626	<b>0.7544890</b>
$\mu$	1.687192	1.367895	1.367732
$\sigma$	0.3833892	0.4231975	0.4307547

After tagging the records in a cluster as in-pattern and out-pattern, the mean of in-pattern records and out-pattern records in each cluster are calculated. Table 6 shows the difference factors of each feature in all three clusters along with the mean and standard deviation of the difference factors of each cluster. The salient features are highlighted in bold. For example, D3 (pain management) and D4 (cleanliness of the hospital) are salient features of cluster 1 while D2 (communication with Nurse) and D5 (quietness) are salient features of cluster 2. The positive values of significant difference factors show high frequency of dichotomous variables and high values for other categorical and continuous variables. Similarly, negative values show a low frequency for dichotomous variables and low values for other categorical and continuous variables. For instance, D3 (pain management) has negative difference factor for cluster one which shows low values of this variable, so the satisfaction with pain management is generally low in this cluster. Similarly, a high satisfaction with pain management can be inferred from cluster three. The salient features of each cluster are presented in Table 7, along with the range of that value or frequency of dichotomous value (High/Low) next to each of them. The salient features with the range of their values are a representation of the cluster, which will be used to create associations.

**Table 6.** Difference Factor of each Feature in each Cluster. Bold numbers are salient features which are out of range of 1 standard deviation from mean.

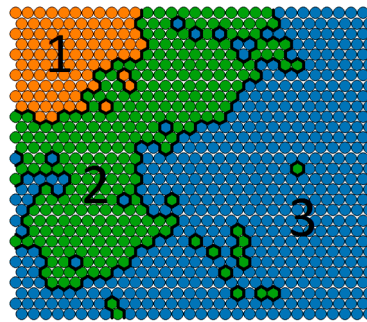
Features	C1	C2	C3
D1	−0.03037039	1.18199255	<b>−1.22503855</b>
D2	0.04416163	<b>−0.20817361</b>	−0.02380192
D3	<b>−1.537553</b>	1.033197	<b>1.041475</b>
D4	<b>−1.554513</b>	1.045382	<b>1.047448</b>
D5	−0.06825377	<b>−0.03673413</b>	−0.03182003
$\mu$	−0.6293057	0.6031327	0.1616524
$\Sigma$	0.7493974	0.5972044	0.8430252

**Table 7.** Salient features.

<i>Cluster</i>	<i>Salient Features</i>	
C1	Pain Management	Low
	Cleanliness	Low
C2	Communication with Nurse	Low
	Quietness	Low
C3	Communication with doctor	Low
	Pain Management	High
	Cleanliness	High

#### 4. Experiment

The methodology proposed in the previous section is implemented in R and is applied to the HCAHPS dataset of a hospital with 2652 records with the same features explained in Section 2.1. First, K-means++ is used to cluster all records based on the patients' responses to the satisfaction questions. A Self-organizing feature map is used to visualize data distribution in each cluster (Figure 2).

**Figure 2.** The clusters produced based on patients' responses to satisfaction questions.

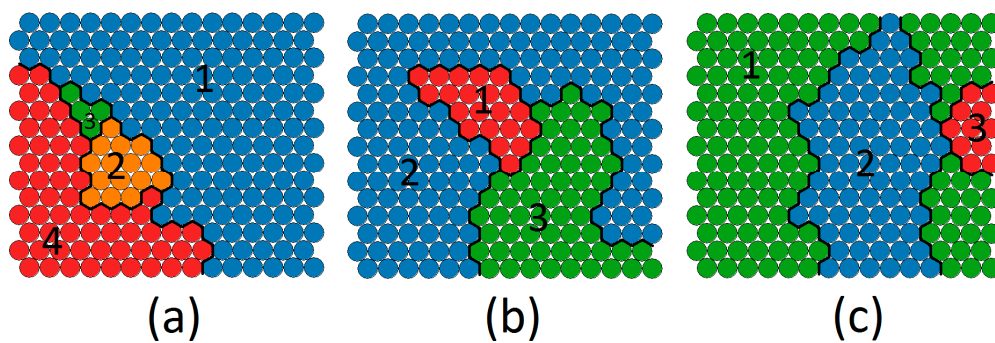
The map shows that Kmeans++ divided data according to the Calinski-Harabasz [15] criterion into three clusters based on their responses to the satisfaction questions. These clusters are shown in different colors in Figure 2. In the next step, the salient features of each cluster are extracted to identify the most important features which constitute a cluster. Extracted salient features were interpreted according to nature of each variable whether it is continuing, categorical, or dichotomous. Table 8 asserts the interpretation of the extracted salient features. For example, cluster 1 represent patients who were well-informed about their symptoms after leaving the hospital and expressed high satisfaction with help after discharge and high overall satisfaction.

**Table 8.** Salient features of clusters in the first layer.

<i>Cluster</i>	<i>Salient Features</i>
C1	• High Satisfaction with Help after discharge
	• High Satisfaction with Symptoms info
	• High Overall Health
C2	• High Satisfaction with Help after discharge
	• High Satisfaction with Symptoms info
C3	• Low Satisfaction with Help after discharge
	• Low Satisfaction with Symptoms info
	• Low Overall Health

Although theoretically salient features are expected be different in different clusters, this expectation could be violated if the clusters are of significantly different sizes. For example, in Table 2, C1, and C2 share some features (i.e., high satisfaction with help after discharge and high satisfaction with symptoms info). This is due to the fact that C3 is much bigger than C1 and C2 (Figure 1). Therefore, it has a significant effect on the value of difference factor, highlighting the shared salient features in C1 and C2. Although it might appear that C2 is redundant, C1 and C2 were divided into two distinct clusters mainly because of their difference in the overall health feature.

The records in each cluster are fed into the second layer of clustering. In this layer, patients are clustered based on their demographic data such as (e.g., age, sex, race, etc.). The first cluster is divided into four demographic sub-clusters and the second and third clusters are both divided into three demographic sub-clusters. The process is outlined in Algorithm 3 and the outcome of this step is visualized in Figure 3.



**Figure 3.** Sub-clusters produced based on patient demographic data (a) demographic sub-clusters of the first satisfaction cluster, (b) demographic sub-clusters of the second satisfaction cluster, (c) demographic subclusters of the third satisfaction cluster.

---

**Algorithm 3. Recommendation extraction.**

---

```

Find Recommendations (data, split_index)
    Satisfaction_questions_data ← data [:] [0:split_index]
    noc ← FindNumberOfClusters (Satisfaction_questions_data )
    Sq_clustered_data = kmeans (Satisfaction_questions_data)
    Sq_salient_dim ← FindingSalientFeatures (noc, Sq_clustered_data, z)
    For i ← 0 to noc
        For j in Sq_clustered_data[i]
            Demographic_data ← data[j] [split_index: number_of_columns(data)]
            D_clustered_data ← kmeans (Demographic_data)
            D_salient_dim ← FindingSalientFeatures (noc, D_clustered_data, z)
            For sq_salient in Sq_salient_dim[i]:
                If sq_salient != 0:
                    For d_salient in D_salient_dim:
                        If d_salient != 0:
                            Recommendation
                                [(col_name(sq_salient), sq_salient)].
                                Add ((col_name(d_salient), d_salient))
    Return recommendation

```

---

Once again salient feature extraction algorithm is applied on each sub-cluster. The results are asserted in Table 9.

Putting the salient features of a satisfaction cluster and the salient features of its demographic sub-cluster, we derived the associations listed in Table 10. These associations represent a possible source of significant satisfaction questions. For instance, the source of low satisfaction with help after discharge according to C3 is communication with old people referred from a physician or Spanish speaking patients admitted from the emergency room. Transforming these associations into recommendations can offer hospital policy changes in favor of both patients and hospitals. The proper recommendation in this case can be putting more effort in communication with elderly people and Spanish people, or hiring a Spanish speaking nurse or interpreter.

It is worth noting, that the method explained in this section, works better in terms of performance and accuracy than in methods such as association rule mining, which has its well-known limitations including a brute-force method for extracting association rules and the risk of finding many irrelevant rules.

**Table 9.** Salient features of all satisfaction clusters and their demographic sub-clusters (Numbers in parenthesis demonstrate the clusters' populations).

Cluster		C1 (1210 Observations)			
Cluster Salient Features	<ul style="list-style-type: none"><li>• High Satisfaction with Help after discharge</li><li>• High Satisfaction with Symptoms info</li><li>• High Overall Health</li></ul>				
Sub-cluster	SC1 (378 obs.)	SC2 (236 obs.)	SC3 (268 obs.)	SC4 (328)	
Sub-cluster Salient Features	<ul style="list-style-type: none"><li>• Mostly not Spanish/Hispanic/Latino</li><li>• Ethnicity</li><li>• Mostly Spanish Language</li></ul>	<ul style="list-style-type: none"><li>• Mostly Old</li><li>• Mostly Emergency Room Admission Source</li><li>• Mostly Medical Principal Reason of Admission</li></ul>	<ul style="list-style-type: none"><li>• Mostly Young</li><li>• Mostly Female</li><li>• Mostly Physician Referral Admission Source</li><li>• Mostly Obstetric Principal Reason of Admission</li></ul>	<ul style="list-style-type: none"><li>• Mostly Old</li><li>• Mostly Physician Referral Admission Source</li><li>• Mostly Surgical Principal Reason of Admission</li></ul>	
Cluster		C2 (602 Observations)			
Cluster Salient Features	<ul style="list-style-type: none"><li>• High Satisfaction with Help after discharge</li><li>• High Satisfaction with Symptoms info</li></ul>				
Sub-cluster	SC1 (141 obs.)	SC2 (224 obs.)	SC3 (237 obs.)		
Sub-cluster Salient Features	<ul style="list-style-type: none"><li>• Mostly Old</li><li>• Mostly Emergency Room Admission Source</li><li>• Mostly Medical Principal Reason of Admission</li></ul>	<ul style="list-style-type: none"><li>• Mostly Physician Referral Admission Source</li><li>• Rarely Medical Principal Reason of Admission</li></ul>	<ul style="list-style-type: none"><li>• Rarely not Spanish/Hispanic/Latino Ethnicity</li><li>• Mostly Spanish Language</li></ul>		
Cluster		C3 (840 Observations)			
Cluster Salient Features	<ul style="list-style-type: none"><li>• Low Satisfaction with Help after discharge</li><li>• Low Satisfaction with Symptoms info</li><li>• Low Overall Health</li></ul>				
Sub-cluster	SC1 (278 obs.)	SC2 (296 obs.)	SC3 (266 obs.)		
Sub-cluster Salient Features	<ul style="list-style-type: none"><li>• Mostly Old</li><li>• Mostly Emergency Room Admission Source</li><li>• Mostly Medical Principal Reason of Admission</li></ul>	<ul style="list-style-type: none"><li>• Rarely not Spanish/Hispanic/Latino Ethnicity</li><li>• Mostly Spanish Language</li></ul>	<ul style="list-style-type: none"><li>• Mostly Physician Referral Admission Source</li><li>• Mostly Obstetric Principal Reason of Admission</li></ul>		

**Table 10.** Derived Associations.

1	Patients who have high Satisfaction with Help After Discharge, have these qualities: Mostly Physician Referral Admission Source, Mostly not Medical Principal Reason of Admission
2	Patients who have high Satisfaction with Symptoms Info, have these qualities: Mostly Physician Referral Admission Source, Rarely Medical Principal Reason of Admission
3	Patients who have high Satisfaction with Help After Discharge, have these qualities: Rarely White Race
4	Patients who have high Satisfaction with Symptoms Info, have these qualities: Rarely White Race

Table 10. Cont.

5	Patients who have high Satisfaction with Help After Discharge, have these qualities: Mostly Old, Mostly Emergency Room Admission Source, Mostly Medical Principal Reason of Admission
6	Patients who have high Satisfaction with Symptoms Info, have these qualities: Mostly Old, Mostly Emergency Room Admission Source, Mostly Medical Principal Reason of Admission
7	Patients who have low Satisfaction with Help After Discharge, have these qualities: Mostly Old, Mostly Physician Referral Admission Source, Mostly Surgical Principal Reason of Admission
8	Patients who have low Satisfaction with Symptoms Info, have these qualities: Mostly Old, Mostly Physician Referral Admission Source, Mostly Surgical Principal Reason of Admission
9	Patients who have low Satisfaction with Help After Discharge, have these qualities: Mostly Young, Mostly Female, Mostly Physician Referral Admission Source, Mostly Obstetric Principal Reason of Admission
10	Patients who have low Satisfaction with Symptoms Info, have these qualities: Mostly Young, Mostly Female, Mostly Physician Referral Admission Source, Mostly Obstetric Principal Reason of Admission
11	Patients who have low Satisfaction with Help After Discharge, have these qualities: Mostly Old, Mostly Emergency Room Admission Source, Mostly Medical Principal Reason of Admission
12	Patients who have low Satisfaction with Symptoms Info, have these qualities: Mostly Old, Mostly Emergency Room Admission Source, Mostly Medical Principal Reason of Admission
13	Patients who have high Satisfaction with Help After Discharge, have these qualities: Mostly Physician Referral Admission Source, Mostly Obstetric Principal Reason of Admission
14	Patients who have high Satisfaction with Symptoms Info, have these qualities: Mostly Physician Referral Admission Source, Mostly Obstetric Principal Reason of Admission
15	Patients who have high Satisfaction with Help After Discharge, have these qualities: Rarely not Spanish/Hispanic/Latino Ethnicity, Rarely English Language
16	Patients who have high Satisfaction with Help After Discharge, have these qualities: Rarely not Spanish/Hispanic/Latino Ethnicity, Rarely English Language
17	Patients who have high Satisfaction with Help After Discharge, have these qualities: Mostly Old, Mostly Emergency Room Admission Source
18	Patients who have high Satisfaction with Symptoms Info, have these qualities: Mostly Old, Mostly Emergency Room Admission Source
19	Patients who have high Overall health, have these qualities: Mostly Old, Mostly Emergency Room Admission Source

## 5. Validation

The associations derived from the two-layer clustering must be validated through standard statistical tests to ensure that they did not occur by chance. This is a very important step for generating reliable associations. Each of the associations is considered as a hypothesis and it is tested based on the whole data set. As described, almost all of the features in the data set are categorical or even dichotomous, which are of multiple groups of studies with unequal sample sizes. In order to work with such data, chi-square test of independence is used for hypothesis testing.

The Chi-square statistic is a non-parametric tool designed to analyze group differences when the dependent variable is measured at a nominal level [17]. This test is robust to the distribution of data. The null hypothesis is stated as  $H_0$ : the two classifications are independent, while the alternative hypothesis is  $H_1$ : the classifications are dependent. The significance of the test is calculated according to the frequency contingency table of the independent classes (Each item in data set belongs only to one class). This value is compared with the critical values in the chi-square table, and if it is larger than this critical value, then the null hypothesis is rejected. Typically, if the chi-square test  $p$ -value is lower than 0.05, the independence of the features in the association is rejected.

Before using chi-square test for validating the associations, we have to make sure that the dataset meets the assumptions of the chi-square. The chi-square test requires that no more than 20% of the frequencies of categories are less than five; otherwise the calculated  $p$ -values may not be accurate. There are many empty or low frequency categories in the patients' contingency table that consists of more than 35% of the frequencies. By combining the columns or rows this issue can be resolved. For example, Table 11 illustrates a 3-way contingency table for testing the 16 association in Table 10. The rows are the satisfaction of symptoms info and the columns are a combination of not Spanish, Hispanic or Latino and the Spanish language. The two-digit binary values in the first row indicate a combination of not Spanish, Hispanic, or Latino Ethnicity (Yes/No) and Spanish Language (Yes/No). For example, 0.0 in the first row indicates that the patients are rarely not Spanish, Hispanic or Latino Ethnicity, and they speak Spanish. In order to get higher frequencies, we can combine the first two rows into a single row to indicate a "low" satisfaction. Similarly, we can merge the last two rows to indicate a "high" satisfaction (Table 12). The rows must be combined in a way to yield an interpretable result. The algorithm proposed for this modification to chi-square test is delineated in Algorithm 4.

**Table 11.** Contingency Table for Satisfaction with Symptoms Info and rarely not Spanish/Hispanic/Latino and Spanish Language.

Satisfaction with Symptoms Info	Rarely Not Spanish/Hispanic/Latino Ethnicity Spanish Language			
	0.0	1.0	0.1	1.1
1	3	180	1710	249
2	1	32	521	63
3	0	6	94	7
4	0	5	72	5

**Table 12.** Contingency Table with Combined Rows.

Satisfaction with Symptoms Info	Rarely Not Spanish/Hispanic/Latino Ethnicity Spanish Language			
	0.0	1.0	0.1	1.1
Low	4	212	2231	256
High	0	11	166	12

---

**Algorithm 4. Modified chi-square test.**

---

*Modified\_chisquare(interaction\_table):*

*Comment: Adjusting interaction table not to have values less than five.*

*Col*  $\leftarrow$  0 to *num\_col(interaction\_table)*:

*If all(table[, col] < 5):*

*table1*  $\leftarrow$  *table* [, -*col*]

*table*  $\leftarrow$  *table1*

*col*  $\leftarrow$  *col* - 1

*numcol*  $\leftarrow$  *numcol* - 1

*col*  $\leftarrow$  *col* + 1

*return chisq.test(table)*

---

The frequencies can further be improved by combining the first two columns as shown in Table 13.

**Table 13.** Contingency with combined Rows and Columns.

Satisfaction with Symptoms Info	Rarely Not Spanish/Hispanic/Latino Ethnicity Spanish Language		
	X.0	0.1	1.1
Low	216	2231	256
High	11	166	12



The interpretation of the contingency table needs to be done carefully when combining the rows or columns. For example, Table 13 indicates that 216 non-Spanish speaking patients were not satisfied with information provided about their symptoms after discharge. The process of combining the rows and columns continuous until the resulting frequencies satisfy the Chi-square test assumptions.

The chi-square test can be used to prove the validity of an association but it does not give any measure to assess the strength of each valid association. After filtering out the invalid associations using chi-square test, we use odds ratio [18] to measure the strength of each association and to rank them based on their strength. Odds ratio is a measure of association between a condition and an outcome. The odds ratio shows the odds that an outcome will occur in a particular condition, as compared to the odds of the outcome occurring in other conditions. A z-test is used to compare two odds ratios. The significance of the z-test is measured by its *p*-value. If the *p*-value is bigger than 2 or smaller than  $-2$ , then the association is considered significant both in terms of dependency and strength. Odds ratio is not dependent to the sample size and it can provide a basis to compare the strengths of the associations extracted from the two-layer clustering between patients' satisfaction identifiers and their demographics. Using these two-step validations, recommendations created because of noise in the data are removed, which alleviates the overfitting that could have happened in the process.

Most associations extracted from the patient dataset have more than two features. Before computing the odds ratio, the contingency table should be converted to a two by two table. The converted table is made by calculating the frequencies of instances that meet the demographic conditions in the derived association and the ones that don't meet these conditions. Table 14 converts the contingency table in Table 13 to a two by two table for calculating the odds ratio.

**Table 14.** Two by Two Table.

<i>Satisfaction with Symptoms Info</i>	Condition: Rarely Not Spanish/Hispanic/Latino Ethnicity High Spanish Language	
	Meet Condition	Do Not Meet Conditions
Low	2447	256
High	177	12

Having converted the contingency table to a two by two table, the odds ratio of each cell in the table is calculated. A z-test is used to compare the odds ratio of the patients meeting the conditions versus the ones who do not having the same satisfaction level. The significance of the z-test is measured by its *p*-value. If the absolute *p*-value is greater than 2, then the association is considered significant both in terms of dependency and strength. Unlike chi-square test, the *p*-value of the z-test can be used both for omitting weak, irrelevant associations and for ranking them. This pseudo code for the validation function used is depicted in Algorithm 5.

---

**Algorithm 5. Validation function.**

---

```

validate_relationships (recommendation)
  Interaction_table ← interact(recommendation)
  Chisqr ← Modifies_chisquare(Interaction_table)
  If Chisqr. P_value < 0.5:
    For row in interaction_table.rows:
      For col in interaction_table.cols:
        Odds_ratio ← Odds_ratio(interaction_table,row,col)
        If odds_ratio.p_value > 2 or odds_ratio.p_value <  $-2$ :
          Return True
  Return False

```

---

Table 15 lists the  $p$ -values of the chi-square and the z-tests for each association derived in Table 10. Associations with significant values are shown in bold figures. If both  $p$ -values values are significant, than the association is considered accurate and reliable. Otherwise, the association is omitted as there is not enough evidence to support its significance. The list of cleaned associations is ranked based on the  $p$ -value of z-test and presented in Table 16.

**Table 15.** Validation results of extracted associations from the HCAHPS data related to one hospital. (Linked to Table 10 by row number). Bold figures are  $p$ -values in accepted test ranges.

1	<i>Chi-square z-test</i>	X-squared = 28.104 $p$ -value = <b>5.00891912614554</b>	$p$ -value = <b>3.455</b> $\times 10^{-6}$
2	<i>Chi-square z-test</i>	X-squared = 7.084 $p$ -value = 0.0183582852328812	$p$ -value = 0.06927
3	<i>Chi-square z-test</i>	X-squared = 1.5969 $p$ -value = 1.31385601104743	$p$ -value = 0.2063
4	<i>Chi-square z-test</i>	X-squared = 0.50012 $p$ -value = -0.760696016958917	$p$ -value = 0.4794
5	<i>Chi-square z-test</i>	X-squared = 30.14 $p$ -value = - <b>2.9118314808416</b>	$p$ -value = <b>0.001506</b>
6	<i>Chi-square z-test</i>	X-squared = 43.594 $p$ -value = - <b>0.212828628801027</b>	$p$ -value = <b>8.558</b> $\times 10^{-6}$
7	<i>Chi-square z-test</i>	X-squared = 26.86 $p$ -value = -1.97247635563556	$p$ -value = <b>0.004825</b>
8	<i>Chi-square z-test</i>	X-squared = 63.926 $p$ -value = <b>3.59338301942709</b>	$p$ -value = <b>1.715</b> $\times 10^{-9}$
9	<i>Chi-square z-test</i>	X-squared = 25.48 $p$ -value = - <b>3.39514132553282</b>	$p$ -value = <b>0.01271</b>
10	<i>Chi-square z-test</i>	X-squared = 61.28 $p$ -value = - <b>5.48685043800752</b>	$p$ -value = <b>1.315</b> $\times 10^{-8}$
11	<i>Chi-square z-test</i>	X-squared = 30.14 $p$ -value = <b>2.9118314808416</b>	$p$ -value = <b>0.001506</b>
12	<i>Chi-square z-test</i>	X-squared = 43.594 $p$ -value = 0.212828628801027	$p$ -value = <b>8.558</b> $\times 10^{-6}$
13	<i>Chi-square z-test</i>	X-squared = 20.061 $p$ -value = <b>3.50130627065847</b>	$p$ -value = <b>0.0001649</b>
14	<i>Chi-square z-test</i>	X-squared = 43.489 $p$ -value = <b>5.57545804293131</b>	$p$ -value = <b>1.937</b> $\times 10^{-9}$
15	<i>Chi-square z-test</i>	X-squared = 8.8791 $p$ -value = <b>2.45951143035353</b>	$p$ -value = <b>0.03094</b>
16	<i>Chi-square z-test</i>	X-squared = 0.9116 $p$ -value = -0.774093020800148	$p$ -value = 0.8226
17	<i>Chi-square z-test</i>	X-squared = 18.715 $p$ -value = - <b>2.85471982135127</b>	$p$ -value = <b>0.002172</b>
18	<i>Chi-square z-test</i>	X-squared = 27.572 $p$ -value = -1.47076729692434	$p$ -value = <b>4.413</b> $\times 10^{-5}$
19	<i>Chi-square z-test</i>	X-squared = 24.691 $p$ -value = -0.231962167837976	$p$ -value = <b>0.0001599</b>

**Table 16.** The list of cleaned associations ranked based on their odds ratio.

1	Patients who have high Satisfaction with Symptoms Info, have these qualities: Mostly Physician Referral Admission Source, Mostly Obstetric Principal Reason of Admission
	<b>z-test</b> $p\text{-value} = 5.57545804293131$
2	Patients who have low Satisfaction with Symptoms Info, have these qualities: Mostly Young, Mostly Female, Mostly Physician Referral Admission Source, Mostly Obstetric Principal Reason of Admission
	<b>z-test</b> $p\text{-value} = -5.48685043800752$
3	Patients who have high Satisfaction with Help After Discharge, have these qualities: Mostly Physician Referral Admission Source, Rarely Medical Principal Reason of Admission
	<b>z-test</b> $p\text{-value} = 5.00891912614554$
4	Patients who have low Satisfaction with Symptoms Info, have these qualities: Mostly Old, Mostly Physician Referral Admission Source, Mostly Surgical Principal Reason of Admission
	<b>z-test</b> $p\text{-value} = 3.59338301942709$
5	Patients who have high Satisfaction with Help After Discharge, have these qualities: Mostly Physician Referral Admission Source, Mostly Obstetric Principal Reason of Admission
	<b>z-test</b> $p\text{-value} = 3.50130627065847$
6	Patients who have low Satisfaction with Help After Discharge, have these qualities: Mostly Young, Mostly Female, Mostly Physician Referral Admission Source, Mostly Obstetric Principal Reason of Admission
	<b>z-test</b> $p\text{-value} = -3.39514132553282$
7	Patients who have high Satisfaction with Help After Discharge, have these qualities: Mostly Old, Mostly Emergency Room Admission Source, Mostly Medical Principal Reason of Admission
	<b>z-test</b> $p\text{-value} = -2.9118314808416$
8	Patients who have high Satisfaction with Help After Discharge, have these qualities: Mostly Old, Mostly Emergency Room Admission Source
	<b>z-test</b> $p\text{-value} = -2.85471982135127$
9	Patients who have high Satisfaction with Help After Discharge, have these qualities: Rarely not Spanish/Hispanic/Latino Ethnicity, Rarely English Language
	<b>z-test</b> $p\text{-value} = 2.45951143035353$
10	Patients who have high Satisfaction with Symptoms Info, have these qualities: Mostly Old, Mostly Emergency Room Admission Source, Mostly Medical Principal Reason of Admission
	<b>z-test</b> $p\text{-value} = -0.212828628801027$

## 6. Turning Associations into Recommendations for Hospitals

The associations extracted in the previous section can be used in two ways to improve patient experience for various patient groups:

1. The valid association can be simply transformed into a set of general applicable recommendations. For instance, based on the second association in Table 16, the system can make the recommendation that “Young ladies whose admission source is physician referral and their reason of admission is obstetrical, need more information about their symptoms when they are being discharged”. In this approach, one recommendation is generated for each correlation, although, the recommendations which are based on patients’ dissatisfaction are probably more useful than ones which are based on patients’ satisfaction.
2. The associations can be used to produce target-based recommendations. Assume that a patient is being admitted to a hospital. The reception takes the patient’s information and relative recommendations would be popped out. For instance, suppose an old patient with physician referral is being admitted for surgical reason. Based on the 4th correlation in Table 4, a recommendation is shown to the health care provider asserting that this patient needs more information about her/his symptoms. This can be accomplished by a simple rule-based expert system.

## 7. Conclusions

In this work, an unsupervised exploratory data analysis methodology is introduced to discover associations between patients' demographics and their various satisfaction identifiers. Such associations are extracted using a two-layer cluster analysis together with extracting the salient features of each cluster. The associations are validated using statistical tests and are ranked based on their significance. The goal was to use such associations to create a patient satisfaction based the recommendation system for hospitals. The methodology was applied to HCAHP data obtained from CAHPS Database and the generated recommendations were validated using statistical tests. In the presented case study in this work, the proposed methodology, extracted nineteen associations from the HCAHPS dataset of a hospital with 2652 records. Ten associations out of nineteen were validated through statistical methods of chi-squared independence test, and odds ratio z-test, which shows the reliability of the proposed recommendation system.

The proposed recommendation system provides knowledge that may be hidden to an expert analyzing the surveys and rectifies the need for a subject-matter expert. The analysis approach is designed specifically for the format of the standard HCAHPS survey; however, it can be extended to other domains in which customer survey plays an important role.

The future work of this study will focus on three aspects:

1. **More extensive data collection:** Our long-term goal is to assess how the recommendations produced by our system can improve patients' loyalty and result in saving costs and time in the long run. Using the preliminary results outlined in this paper, our goal is to obtain a more comprehensive data set which includes data on whether the patients have come back to the hospital if medical services were needed, and to examine the relationship between customer loyalty and their satisfaction identifiers.
2. **Handling skip questions:** In this study, a single imputation method based on K-nearest neighbor (KNN) is used to impute missing values. Other popular approaches, such as multiple imputations by chained equations (MICE) [19], should be explored in future for imputing both categorical and continues variables. Also, other approaches for handling skip questions should be examined to better distinguish between non-applicable and missing data.
3. **Alternative distance measures for K-means:** In this study, we used Euclidean distance for clustering. Euclidean distance is typically used for continuous data where data are seen as points in the Euclidean space. Since HCAHPS data consist of mixed numeric and categorical variables, we should examine other types of distance measures such as cosine, Jaccard, Overlap, Occurrence Frequency, etc. [20] and compare the quality of recommendations produced by each measure. In addition, since there are two layers of clustering and the intrinsic characteristics of data points in each layer vary, and a different distance function can be used for each layer.

**Acknowledgments:** The CAHPS® data used in this analysis were provided by the CAHPS Database. The CAHPS Database is funded by the U.S. Agency for HealthCare Research and Quality (AHRQ) and administered by Westat under Contract No. HHSA290201300003C.

**Author Contributions:** Shahram Rahimi studied conception and design; Mohammadreza Khoie and Tannaz Sattari Tabrizi and Nina Marhamati acquired data; All authors analysed and interpreted data and evaluated results; Tannaz Sattari Tabrizi and Mohammad Reza Khoie and Elham Sahebkar Khorasani drafted the manuscript; Shahram Rahimi and Elham Sahebkar Khorasani and Nina Marhamati did critical revisions.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A

The HCAHPS survey has been changed through the time. In this appendix A, we brought the questionnaire of March 2012.

## HCAHPS Survey

### SURVEY INSTRUCTIONS

- ◆ You should only fill out this survey if you were the patient during the hospital stay named in the cover letter. Do not fill out this survey if you were not the patient.
- ◆ Answer all the questions by checking the box to the left of your answer.
- ◆ You are sometimes told to skip over some questions in this survey. When this happens you will see an arrow with a note that tells you what question to answer next, like this:

- ☐ Yes  
☒ No → **If No, Go to Question 1**

***You may notice a number on the survey. This number is used to let us know if you returned your survey so we don't have to send you reminders.***

***Please note: Questions 1-22 in this survey are part of a national initiative to measure the quality of care in hospitals. OMB #0938-0981***

Please answer the questions in this survey about your stay at the hospital named on the cover letter. Do not include any other hospital stays in your answers.

#### YOUR CARE FROM NURSES

- |   |  |
|---|--|
| <p><b>1. During this hospital stay, how often did nurses treat you with <u>courtesy and respect</u>?</b></p> <p> <sup>1</sup> <input type="checkbox"/> Never<br/> <sup>2</sup> <input type="checkbox"/> Sometimes<br/> <sup>3</sup> <input type="checkbox"/> Usually<br/> <sup>4</sup> <input type="checkbox"/> Always         </p> <p><b>2. During this hospital stay, how often did nurses <u>listen carefully to you</u>?</b></p> <p> <sup>1</sup> <input type="checkbox"/> Never<br/> <sup>2</sup> <input type="checkbox"/> Sometimes<br/> <sup>3</sup> <input type="checkbox"/> Usually<br/> <sup>4</sup> <input type="checkbox"/> Always         </p> | <p><b>3. During this hospital stay, how often did nurses <u>explain things</u> in a way you could understand?</b></p> <p> <sup>1</sup> <input type="checkbox"/> Never<br/> <sup>2</sup> <input type="checkbox"/> Sometimes<br/> <sup>3</sup> <input type="checkbox"/> Usually<br/> <sup>4</sup> <input type="checkbox"/> Always         </p> <p><b>4. During this hospital stay, after you pressed the call button, how often did you get help as soon as you wanted it?</b></p> <p> <sup>1</sup> <input type="checkbox"/> Never<br/> <sup>2</sup> <input type="checkbox"/> Sometimes<br/> <sup>3</sup> <input type="checkbox"/> Usually<br/> <sup>4</sup> <input type="checkbox"/> Always<br/> <sup>9</sup> <input type="checkbox"/> I never pressed the call button         </p> |
|---|--|

**YOUR CARE FROM DOCTORS**

5. During this hospital stay, how often did doctors treat you with courtesy and respect?

<sup>1</sup> ☐ Never  
<sup>2</sup> ☐ Sometimes  
<sup>3</sup> ☐ Usually  
<sup>4</sup> ☐ Always

6. During this hospital stay, how often did doctors listen carefully to you?

<sup>1</sup> ☐ Never  
<sup>2</sup> ☐ Sometimes  
<sup>3</sup> ☐ Usually  
<sup>4</sup> ☐ Always

7. During this hospital stay, how often did doctors explain things in a way you could understand?

<sup>1</sup> ☐ Never  
<sup>2</sup> ☐ Sometimes  
<sup>3</sup> ☐ Usually  
<sup>4</sup> ☐ Always

**THE HOSPITAL ENVIRONMENT**

8. During this hospital stay, how often were your room and bathroom kept clean?

<sup>1</sup> ☐ Never  
<sup>2</sup> ☐ Sometimes  
<sup>3</sup> ☐ Usually  
<sup>4</sup> ☐ Always

9. During this hospital stay, how often was the area around your room quiet at night?

<sup>1</sup> ☐ Never  
<sup>2</sup> ☐ Sometimes  
<sup>3</sup> ☐ Usually  
<sup>4</sup> ☐ Always

**YOUR EXPERIENCES IN THIS HOSPITAL**

10. During this hospital stay, did you need help from nurses or other hospital staff in getting to the bathroom or in using a bedpan?

<sup>1</sup> ☐ Yes  
<sup>2</sup> ☐ No → If No, Go to Question 12

11. How often did you get help in getting to the bathroom or in using a bedpan as soon as you wanted?

<sup>1</sup> ☐ Never  
<sup>2</sup> ☐ Sometimes  
<sup>3</sup> ☐ Usually  
<sup>4</sup> ☐ Always

12. During this hospital stay, did you need medicine for pain?

<sup>1</sup> ☐ Yes  
<sup>2</sup> ☐ No → If No, Go to Question 15

13. During this hospital stay, how often was your pain well controlled?

<sup>1</sup> ☐ Never  
<sup>2</sup> ☐ Sometimes  
<sup>3</sup> ☐ Usually  
<sup>4</sup> ☐ Always

14. During this hospital stay, how often did the hospital staff do everything they could to help you with your pain?

<sup>1</sup> ☐ Never  
<sup>2</sup> ☐ Sometimes  
<sup>3</sup> ☐ Usually  
<sup>4</sup> ☐ Always

15. During this hospital stay, were you given any medicine that you had not taken before?

<sup>1</sup> ☐ Yes

<sup>2</sup> ☐ No → If No, Go to Question 18

16. Before giving you any new medicine, how often did hospital staff tell you what the medicine was for?

<sup>1</sup> ☐ Never

<sup>2</sup> ☐ Sometimes

<sup>3</sup> ☐ Usually

<sup>4</sup> ☐ Always

17. Before giving you any new medicine, how often did hospital staff describe possible side effects in a way you could understand?

<sup>1</sup> ☐ Never

<sup>2</sup> ☐ Sometimes

<sup>3</sup> ☐ Usually

<sup>4</sup> ☐ Always

#### WHEN YOU LEFT THE HOSPITAL

18. After you left the hospital, did you go directly to your own home, to someone else's home, or to another health facility?

<sup>1</sup> ☐ Own home

<sup>2</sup> ☐ Someone else's home

<sup>3</sup> ☐ Another health facility → If Another, Go to Question 21

19. During this hospital stay, did doctors, nurses or other hospital staff talk with you about whether you would have the help you needed when you left the hospital?

<sup>1</sup> ☐ Yes

<sup>2</sup> ☐ No

20. During this hospital stay, did you get information in writing about what symptoms or health problems to look out for after you left the hospital?

<sup>1</sup> ☐ Yes

<sup>2</sup> ☐ No

#### OVERALL RATING OF HOSPITAL

Please answer the following questions about your stay at the hospital named on the cover letter. Do not include any other hospital stays in your answers.

21. Using any number from 0 to 10, where 0 is the worst hospital possible and 10 is the best hospital possible, what number would you use to rate this hospital during your stay?

<sup>0</sup> ☐ 0 Worst hospital possible

<sup>1</sup> ☐ 1

<sup>2</sup> ☐ 2

<sup>3</sup> ☐ 3

<sup>4</sup> ☐ 4

<sup>5</sup> ☐ 5

<sup>6</sup> ☐ 6

<sup>7</sup> ☐ 7

<sup>8</sup> ☐ 8

<sup>9</sup> ☐ 9

<sup>10</sup> ☐ 10 Best hospital possible



**22. Would you recommend this hospital to your friends and family?**

- <sup>1</sup> ☐ Definitely no  
<sup>2</sup> ☐ Probably no  
<sup>3</sup> ☐ Probably yes  
<sup>4</sup> ☐ Definitely yes

**ABOUT YOU**


---

There are only a few remaining items left.

**23. In general, how would you rate your overall health?**

- <sup>1</sup> ☐ Excellent  
<sup>2</sup> ☐ Very good  
<sup>3</sup> ☐ Good  
<sup>4</sup> ☐ Fair  
<sup>5</sup> ☐ Poor

**24. What is the highest grade or level of school that you have completed?**

- <sup>1</sup> ☐ 8th grade or less  
<sup>2</sup> ☐ Some high school, but did not graduate  
<sup>3</sup> ☐ High school graduate or GED  
<sup>4</sup> ☐ Some college or 2-year degree  
<sup>5</sup> ☐ 4-year college graduate  
<sup>6</sup> ☐ More than 4-year college degree

**25. Are you of Spanish, Hispanic or Latino origin or descent?**

- <sup>1</sup> ☐ No, not Spanish/Hispanic/Latino  
<sup>2</sup> ☐ Yes, Puerto Rican  
<sup>3</sup> ☐ Yes, Mexican, Mexican American, Chicano  
<sup>4</sup> ☐ Yes, Cuban  
<sup>5</sup> ☐ Yes, other Spanish/Hispanic/Latino

**26. What is your race? Please choose one or more.**

- <sup>1</sup> ☐ White  
<sup>2</sup> ☐ Black or African American  
<sup>3</sup> ☐ Asian  
<sup>4</sup> ☐ Native Hawaiian or other Pacific Islander  
<sup>5</sup> ☐ American Indian or Alaska Native

**27. What language do you mainly speak at home?**

- <sup>1</sup> ☐ English  
<sup>2</sup> ☐ Spanish  
<sup>3</sup> ☐ Chinese  
<sup>4</sup> ☐ Russian  
<sup>5</sup> ☐ Vietnamese  
<sup>6</sup> ☐ Some other language (please print): \_\_\_\_\_

**THANK YOU**

**Please return the completed survey in the postage-paid envelope.**

**References**

1. Lehtonen, R.; Pahkinen, E. *Practical Methods for Design and Analysis of Complex Surveys*; John Wiley & Sons: Hoboken, NJ, USA, 2004.
2. Kenett, R.; Salini, S. *Modern Analysis of Customer Surveys: With Applications Using R (Vol. 117)*; John Wiley & Sons: Hoboken, NJ, USA, 2011.
3. Giordano, L.A.; Elliott, M.N.; Goldstein, E.; Lehrman, W.G.; Spencer, P.A. Development, implementation, and public reporting of the HCAHPS survey. *Med. Care Res. Rev.* **2010**, *67*, 27–37. [[CrossRef](#)] [[PubMed](#)]
4. Stratford, N.J. Patient perception of pain care in the United States: A 5-year comparative analysis of hospital consumer assessment of health care providers and systems. *Pain Phys.* **2014**, *17*, 369–377.
5. Sheetz, K.H.; Seth, A.W.; Micah, E.G.; Darrell, A.C., Jr.; Michael, J.E. Patients' perspectives of care and surgical outcomes in Michigan: an analysis using the CAHPS hospital survey. *Ann. Surg.* **2014**, *260*, 5–9. [[CrossRef](#)] [[PubMed](#)]
6. Goldstein, E.; Marc, N.E.; William, G.L.; Katrin, H.; Laura, A.G. Racial/ethnic differences in patients' perceptions of inpatient care using the HCAHPS survey. *Med. Care Res. Rev.* **2010**, *67*, 74–92. [[CrossRef](#)] [[PubMed](#)]
7. Elliott, M.N.; William, G.L.; Megan, K.B.; Elizabeth, G.; Katrin, H.; Laura, A.G. Gender differences in patients' perceptions of inpatient care. *Health Serv. Res.* **2012**, *47*, 1482–1501. [[CrossRef](#)] [[PubMed](#)]

8. Elliott, M.N.; William, G.L.; Elizabeth, G.; Katrin, H.; Megan, K.B.; Laura, A.G. Do hospitals rank differently on HCAHPS for different patient subgroups? *Med. Care Res. Rev.* **2010**, *67*, 56–73. [[CrossRef](#)] [[PubMed](#)]
9. Klinkenberg, W.; Dean, S.B.; Brian, M.W.; Koichiro, O.; Joe, M.I.; Jan, C.G.; Wm Claiborne, D. Inpatients' willingness to recommend: A multilevel analysis. *Health Care Manag. Rev.* **2011**, *36*, 349–358. [[CrossRef](#)] [[PubMed](#)]
10. Kleinbaum, D.; Lawrence, K.; Azhar, N.; Eli, R. *Applied Regression Analysis and Other Multivariable Methods*; Nelson Education: Scarborough, ON, Canada, 2013.
11. Scholkopf, B.; Klaus-Robert, M. Fisher Discriminant Analysis with Kernels. In Proceedings of the Neural Networks for Signal Processing IX, Madison, WI, USA, 25 August 1999.
12. Thurstone, L.L. Multiple factor analysis. *Psychol. Rev.* **1931**, *38*, 406. [[CrossRef](#)]
13. Batista, G.E.; Monard, M.C. A Study of K-Nearest Neighbour as an Imputation Method. *HIS* **2002**, *87*, 251–260.
14. Ailon, N.; Ragesh, J.; Claire, M. Streaming  $k$ -means approximation. In *Advances in Neural Information Processing Systems*; Columbia University: New York, NY, USA, 2009; pp. 10–18.
15. Caliński, T.; Jerzy, H. A dendrite method for cluster analysis. *Commun. Stat.-Theory Methods* **1974**, *3*, 1–27. [[CrossRef](#)]
16. Azcarraga, A.P.; Hsieh, M.H.; Pan, S.L.; Setiono, R. Extracting salient dimensions for automatic SOM labeling. *IEEE Trans. Syst. Man Cybern. Part C* **2005**, *35*, 595–600. [[CrossRef](#)]
17. McHugh, M.L. The chi-square test of independence. *Biochem. Med.* **2013**, *23*, 143–149. [[CrossRef](#)]
18. Bland, J.M.; Douglas, G.A. The odds ratio. *BMJ* **2000**, *320*, 1468. [[CrossRef](#)] [[PubMed](#)]
19. White, I.R.; Patrick, R.; Angela, M.W. Multiple imputation using chained equations: Issues and guidance for practice. *Stat. Med.* **2011**, *30*, 377–399. [[CrossRef](#)] [[PubMed](#)]
20. Boriah, S.; Varun, C.; Vipin, K. Similarity measures for categorical data: A comparative evaluation. In Proceedings of the 2008 Society for Industrial and Applied Mathematics (SIAM) International Conference on Data Mining, Atlanta, GA, USA, 24–26 April 2008; pp. 243–254.



© 2017 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).