**Related Slide Deck Summary- PCA Example**

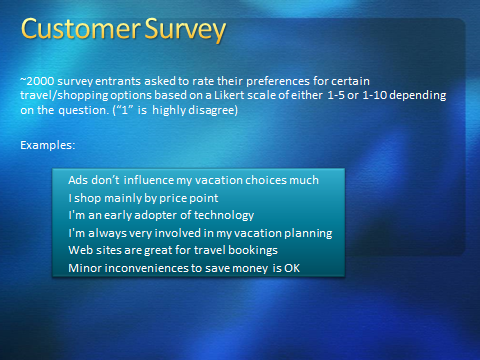
**Overview**

The goal of the Principal Component Analysis application described below was twofold.

The first goal was to reduce the dimensionality of an example customer survey using PCA and the second goal was to concurrently familiarize myself with the R language.

**Task**

The task at hand was to apply PCA in order to reduce dimensionality of a customer survey with the ultimate goal of segmenting the customers. The survey consisted of 26 questions pertaining to customer shopping/travel preferences measured on a Likert scale of either 1 through 5 or 1 through 10 depending on the question. Sample questions that capture the general idea of the survey appear below. We aimed to compress the information from the survey into a smaller number of final questions that we can use but which still capture the majority of the information inherent in the entire set.



**Final Results**

Ultimately, we were able to compress the twenty-six original questions to just the eight questions shown below, which also suggest a customer segmentation we may be able to apply to an advertising model.



For example, each of the four question groups in the above slide appears to highlight specific customer purchasing behaviors. There are four segments including,

**luxury** (q2\_1 and q2\_3),

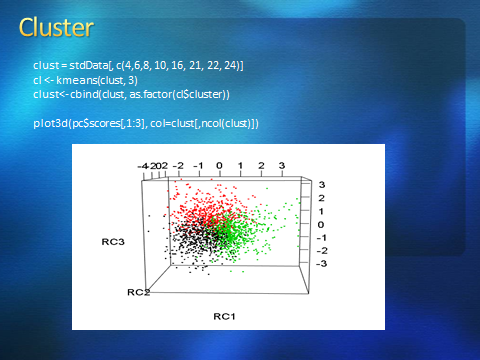
**economy** (q2\_17 and q2\_18),

**impulse buyers** (q1\_5 and q1\_7),

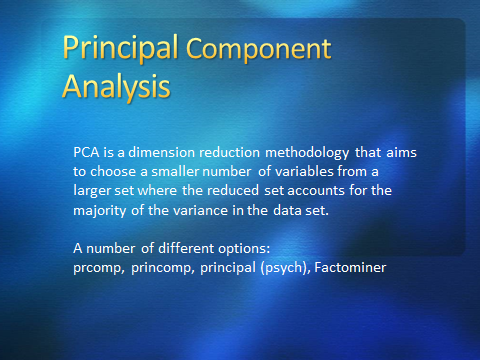
**not convinced of web related purchasing** (q2\_10 and q2\_2).

Each segment could have marketing campaigns developed specifically for that segment. Luxury travel groups would clearly receive different ads than economy groups. Impulse buyers may be targeted with more exploding offer type ads, while those who eschew the web might be candidates for phone follow ups rather than email or other types of contacts.

The segments can be visualized to get a perspective on relationships and relative size of each segment and how much separation exists. In the example below, only three of the segments are displayed but the graph illustrates the point.

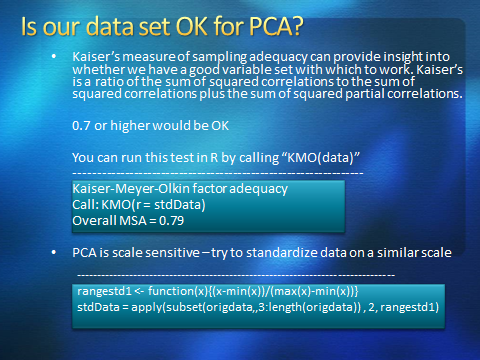


**Methodology**



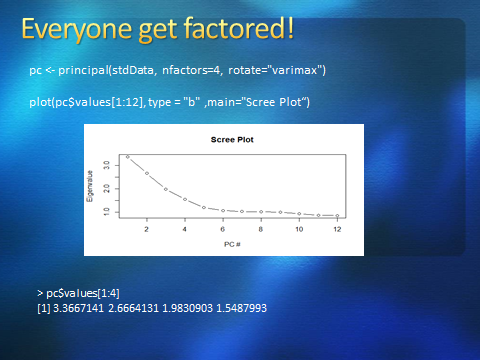
I reviewed at least four of the major PCA options provided by R including the functions ***prcomp*** and ***princomp***, as well as two packages named ***psych*** and ***Factominer***. All are feasible choices, but the ***psych*** package is very comprehensive and easy to use. ***Factominer*** supports a visual interface but was overkill for the task at hand. One of the differences among the libraries is that some (like ***prcomp***) use singular value decomposition while others may employ an eigenvalue decomposition approach to acquire a solution

After choosing the right PCA generator, one of the first questions to answer will be “*Is our data set OK for PCA?”* We would like to ensure that we have sufficient correlation in the data to actually make use of dimension reduction. One approach is to utilize Kaiser’s measure which returns a value related to the correlation. A value of 0.70 that we obtained is generally OK for consideration.

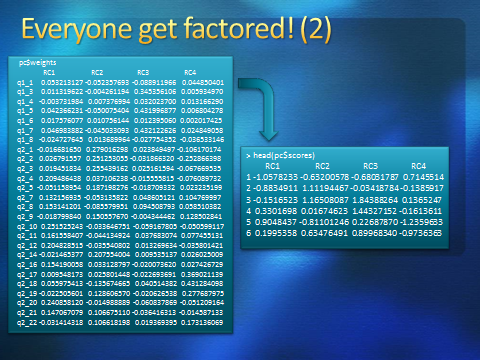


Since the PCA methodology is sensitive to variance, standardizing the data set is an important step. The code used to obtain the KMO result as well as to standardize the data is as set forth in the slide above. After the data set has been prepared, we can utilize the ***principal*** function in the ***psych*** package to create our components.

Invoking the ***principal*** function in the ***psych*** package is illustrated in the following slide. The call shows a request for four principal components using varimax rotation. The scree plot suggests 4 or 5 components capture most of the variance. The associated eigenvalues are displayed from the ***pc*** data structure created by the call to ***principal.*** The eigenvaluesrepresent the magnitude of the variance captured by the associated eigenvector. The eigenvectors represent the direction or axis of variance. They are orthogonal and uncorrelated by design.



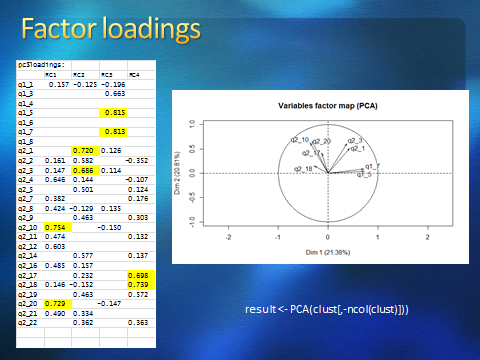
Each PC is a linear combination of the original features in the data set. We can inspect how such information was mapped into our four new PCs by examining the ***pc$weights*** and ***pc$scores*** structures. The relationships are as shown in the slide below. One can see that each of the PCs under ***pc$weights*** is a combination of each of the original questions. Further, data has also been mapped into the PC space in the ***pc$scores*** structure.



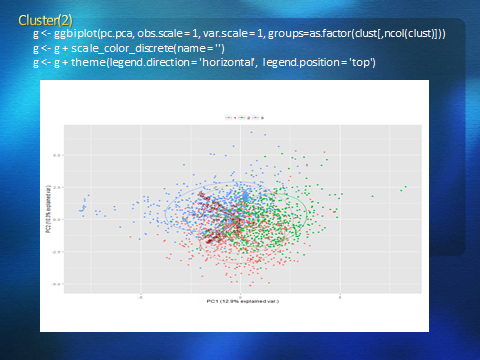
The original twenty six questions have now been reduced to the information in the eight questions shown below.



The related factor loadings for the PCs (contained in the ***pc$loadings*** structure) are depicted in the following slide. I chose to use loadings ~ .7 or greater as they made the interpretation of the loadings easier. PC1 can be interpreted as the segment of customers “unlikely to use the web” based on the questions q2\_10 and q2\_20, while PC2 is the luxury segment, PC3 is the impulse shoppers and PC4 is the economy travel group. The squared loadings are related to the amount of variance of each variable represented in a PC.



One can cluster on these segments given demographic information that may help to identify particular customer characteristics like gender, age or geographic location. There are numerous ways to illustrate such clusters (like the 3d plot earlier) and also a convenient package called ***ggbiplot*** which is demonstrated below. Three of the clusters are represented here for illustration purposes and show the versatility of the plotting capabilities.



**Wrap Up**

R provides numerous and flexible processing for PCA analysis. We demonstrated using the function ***principal*** from the ***psych*** package to extract a maximum amount of information from a survey questionnaire. We also utilized two different plotting utilities (***plot3d*** and ***ggbiplot***) available for illustration purposes. K-means clustering was applied to the data for further analysis after PCA. Ultimately, we were able to show a possible business application to find target segments for a marketing campaign.

