Review of Prior Research

We are working with set of data collected in 1976 from surveys on how people use their time. In previous work, we examined the data set via principal component analysis (PCA) and factor analysis (FA). Assessment of the correlation matrix for the time-use values in the data, reveals multi-collinearity. We can account for 80% of the variation in the predictor variables with the first 3 principal components. PC1 is negatively loaded for Transport and Professional, while being positively loaded for "Homemaking" and "Child Care". PC2 is positively loaded for "Shopping" and "Personal", while PC3 is negatively loaded for "TV". Factor analysis did not yield additional insights into the data set beyond what was found by PCA.

Distance Measures and Input Matrices

Using the silhouette coefficient as a way to evaluate the distances between clusters, we can see that using 10 clusters for the time-use data provides the best separation between clusters.

Table 1 – Time-use Clusters

```
nclusters: 2, silhouette coefficient: 0.341060
nclusters: 3, silhouette coefficient: 0.270682
nclusters: 4, silhouette coefficient: 0.300029
nclusters: 5, silhouette coefficient: 0.345337
nclusters: 6, silhouette coefficient: 0.366910
nclusters: 7, silhouette coefficient: 0.363074
nclusters: 8, silhouette coefficient: 0.396322
nclusters: 9, silhouette coefficient: 0.408737
nclusters: 10, silhouette coefficient: 0.437558
highest silhouette score is 0.437558 for nclusters = 10
```

Based on these calculations, I am using k=10 for my number of clusters on the time-use data.

Performing a similar calculation for the demographic data, reveals that 2 is an optimum number of clusters to maximize distance between clusters.

Table 2 – Demographic Clusters

```
nclusters: 2, silhouette coefficient: 0.565575
nclusters: 3, silhouette coefficient: 0.502753
nclusters: 4, silhouette coefficient: 0.471939
nclusters: 5, silhouette coefficient: 0.407075

highest silhouette score is 0.565575 for nclusters = 2
```

The silhouette measure, which uses Euclidean distance as part of its calculations, is much easier to interpret than the raw Euclidean distances between points when trying to determine a value

for the number of clusters to use. A short sample of the Euclidean computations is shown below to illustrate the complexity of using the distance directly to select number of clusters:

Table 3 – Sample of Euclidean Distances between Time-use observations

	0	1	2	3	4	5	6	\
0	0.000000	3.343466	6.399161	0.540076	5.006690	2.016541	3.736413	
1	3.343466	0.000000	5.017914	3.715235	3.318706	2.667792	1.217609	
2	6.399161	5.017914	0.000000	6.588570	1.758347	5.884904	5.372445	
3	0.540076	3.715235	6.588570	0.000000	5.242857	2.444840	4.174042	
4	5.006690	3.318706	1.758347	5.242857	0.000000	4.489123	3.815765	

	7	8	9	10	11	12	13	\
0	3.530609	4.077053	6.747309	3.586922	5.944452	3.967678	5.070485	
1	4.843942	3.785023	6.138705	4.938963	5.764742	5.366504	4.531486	
2	7.161573	6.077610	4.092063	7.195958	4.386985	7.416292	5.638723	
3	3.361486	4.026130	6.714099	3.399411	5.880593	3.786174	5.091172	
4	6.051285	4.840508	4.315909	6.097959	4.325177	6.419313	4.870909	

The Sklearn implementation of K-Means does not specify precisely how it is computing distances in the algorithm, Euclidean distance is assumed since that seems to be used everywhere else in the clustering API; per the documentation, Sklearn does use Euclidean distances for calculating hierarchy clusters.

Clusters of Activities

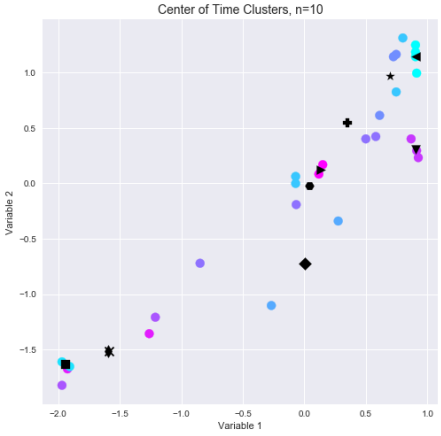
Using the value of k=10 for my number of clusters, as determined from the silhouette metrics shown in Table 1, I generated the clusters shown in Table 4. The end-results of cluster creation, unsurprisingly, generated clusters which exactly mapped to the time-use variables. The order in which the assignments were made differ between the 2 methods as shown in Table 4.

Table 4

cluster membership by variable for K-Means time-use	**cluster membership by variable for Hierarchy time-use**
cluster variable	cluster variable
0 housework	0 leisure
1 professional	1 sleep
2 tv	2 tv
3 sleep	3 mealtime
4 personal	4 childcare
5 leisure	5 personal
6 shopping	6 shopping
7 mealtime	7 transport
8 childcare	8 housework
9 transport	9 professional

In selecting a method to summarize, I elected to go with K-Means. I find it somewhat easier to explain that sort of distance. When K-Means assigns a point into a cluster, it looks at the amount of difference (i.e. distance) between that point and the others in the data set, so that points which are the most similar (closest) are assigned to the same cluster. This has the result of creating clusters with members that are more like each other, than they are like the rest of the sample. A visual example is given in Figure 1. The salient point being the black shapes are the cluster centroids; you can see how each black shape is in the middle of 2 or more points of the same color. The colors represent the different clusters. Looking at the black diamond at 0.0, -.75, you can see how it is the center of the 2 cadet blue dots, and that they are away from other points, forming their own cluster.

Figure 1



Since you set the number of clusters up-front, the final clusters created may not be optimal. For example, if I set the number of cluster to be 4, instead of 10, the cluster assignments would change to:

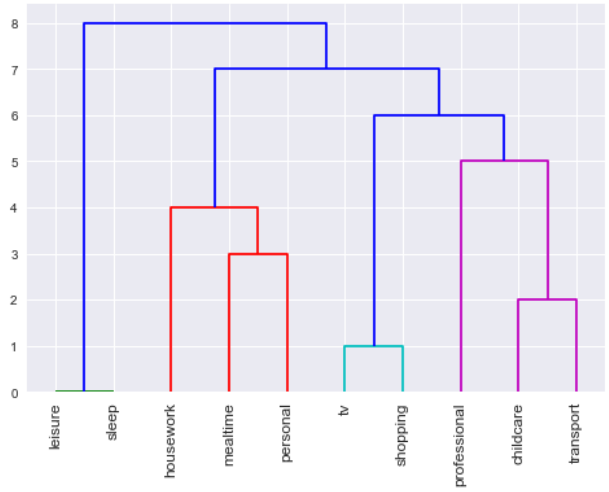
Table 5 – Results for 4 clusters

cluster membership by variable for K-Means time-use, n_clusters = 4	
cluster variable	cluster variable
0 housework	2 professional
0 childcare	2 transport
1 shopping	3 mealtime
1 personal	3 sleep
1 tv	3 leisure

indicating Housework and Childcare are more like each other than they are like Shopping, Personal and TV. I conclude from this, to a large extent, the number of clusters selected depends on both measures like the Silhouette coefficient, and what you are trying to use the data for; do the clusters support a view of the data that addresses your question?

For completeness in explanations, Hierarchical clusters are often visualized as dendrograms, like Figure 2. I do not find them particularly easy to explain. In this case, since there is a 1:1 mapping between the leaf nodes and the clusters, it is a straightforward interpretation, but that is not universally the case.

Figure 2



Clusters of Demographic Groups

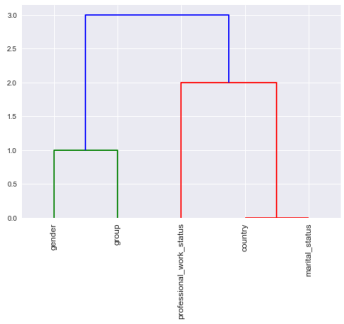
The demographic data is text-based. In order to work with it, I used Pandas Categorical method to convert it to numeric data. Computing the silhouette metrics for 2 to 5 possible clusters (shown in Table 2), yielded 2 as the best number of clusters to specify for this problem. The variables were assigned to clusters as follows:

Table 6

cluster membership by variable
cluster variable
0 gender
0 professional_work_status
0 marital_status
0 country
1 group

In this case, the dendrogram for the Hierarchy solution is particularly confusing. It looks like the colors (red/green) should mean 2 clusters, and the variables in each cluster. However, Group is its own cluster, and gender is assigned to be with the other variables.

Figure 3



Since the Group variable is basically an aggregate code which encapsulates all the other variables, it makes sense that it would be a single cluster and the other variables would make up the second.

Model Comparison and Recommendation

In order to know which solution to propose to management, it would be necessary to understand what consumer groups they are trying to target. If the goal is to market a specific new product to women, then a simple gender-based-demographic clustering solution might be sufficient. If the goal is to market to "commuters" i.e. professionals who spend a fair bit of time doing "transportation", then time-use clustering would be appropriate. On this limited sample set, the outcomes for K-Means and Hierarchy have been the same, so there is no clear preference for which method to use in making the clusters.

In this exercise, thinking about marketing, I would recommend the data pivot that best aligns with the planned actions to be taken as a result of the analysis. If we are marketing a product that is related to how people use time, then pivot on the time variables. If we are marketing a product that relates to a specific activity, drill into those data; if the product has regional appeal, used the demographic data. As with much of analysis, which set of techniques to use and how to manipulate the data is driven by the questions being asked and the planned response to answers.