Introduction

For this assignment, we are looking at data collected on how people spend their time. The data were collected via surveys done in 1976. The population surveyed is drawn from a number of countries. Men and women, married and unmarried, professional and not professional participants are all part of the sample. Participants are divided into 28 groups, depending on their country, gender, marital status and professional status. The data show the average number of minutes spent by each group in 10 different activity categories.

Our task is to look at these data first using principal component analysis, then rotate a subset of the principal components. Next, we will look at the data using factor analysis. Finally, we will compare the models produced.

Review Sample Data and Exploratory Data Analysis

The 10 categories of activity, showing the average use of time, are summarized as:

Table 1

	professional	transport	housework	childcare	shopping
count	28.000000	28.000000	28.000000	28.000000	28.000000
mean	448.857143	86.071429	276.964286	33.321429	108.678571
std	226.976376	48.095529	198.606718	30.457078	32.514445
min	10.000000	0.000000	50.000000	0.000000	52.000000
25%	356.750000	47.500000	96.500000	10.000000	85.000000
50%	535.000000	95.500000	256.000000	22.000000	112.000000
75%	630.750000	127.000000	423.500000	56.000000	131.000000
max	655.000000	148.000000	710.000000	110.000000	170.000000

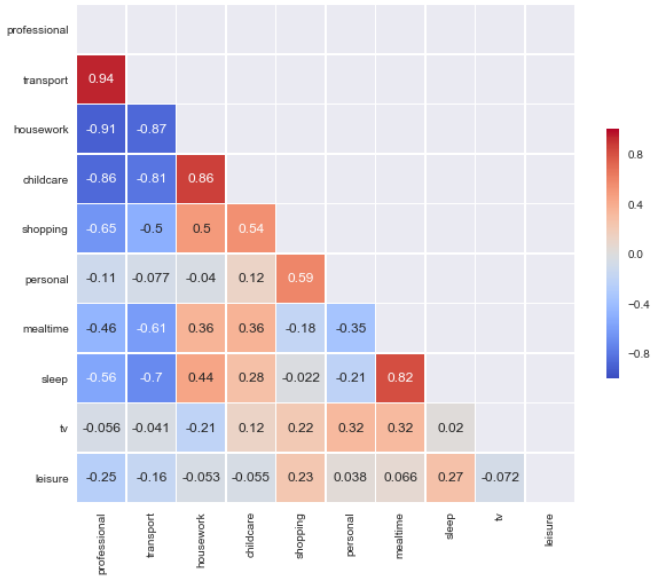
	personal	mealtime	sleep	tv	leisure
count	28.000000	28.000000	28.000000	28.000000	28.000000
mean	94.857143	118.071429	785.607143	99.428571	348.428571
std	11.555708	25.703334	29.586457	39.408994	64.294132
min	77.000000	85.000000	745.000000	40.000000	228.000000
25%	89.500000	100.000000	761.500000	64.750000	308.750000
50%	92.000000	110.000000	775.000000	91.500000	361.000000
75%	96.250000	132.500000	808.250000	122.750000	388.250000
max	130.000000	180.000000	848.000000	180.000000	475.000000

Sleep has the highest average, followed by Professional. Since everyone sleeps it makes sense that it would have the highest mean. Also, since many people work, it makes sense that Professional would have the next highest mean. Professional has the largest standard deviation, indicating a wide range in how many minutes, on average, people spend working

each week. Personal has the smallest standard deviation, indicating there is a small range for how much time people spend on that activity.

The correlation matrix for the time spent on these activities may be visualized as:

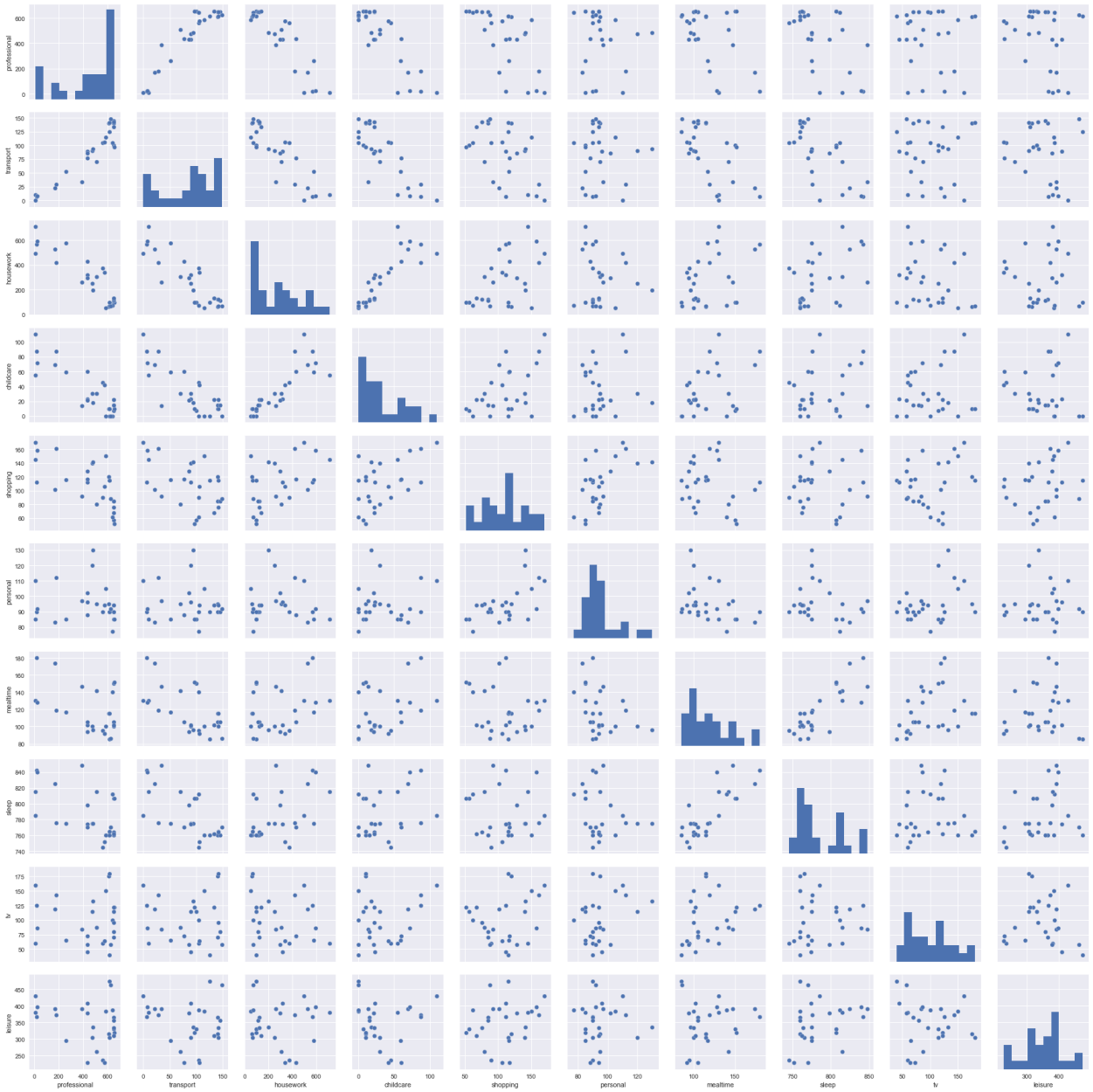
Figure 1



There are both positively and negatively correlated activities, as well as uncorrelated activities.

The scatterplot matrix for time use is shown in Figure 2. We can see that there are a few activities which show some amount of linear correlation. For example, Transport and Professional seem to have a positive linear relationship. For most countries, Childcare and Housework also seem to have a positive relationship; the spread of points in the upper right of that scatterplot is indicative of some per-country variation. Other relationships, like Childcare and Professional seem to have a loose negative correlation.

Figure 2


professional
transport
housework
childcare
shopping
personal
mealtime
sleep
tv
leisure

Principal Component Analysis — Un-rotated

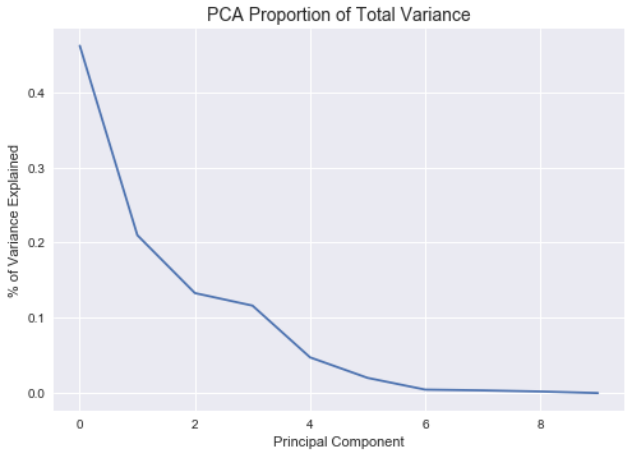
Using all 10 time-use variables, to construct the principal components, we get the following:

Table 2

```
PC 1 accounts for 46.2% of variation; cummulative variation is: 46.2%
PC 2 accounts for 21% of variation; cummulative variation is: 67.2%
PC 3 accounts for 13.3% of variation; cummulative variation is: 80.5%
PC 4 accounts for 11.6% of variation; cummulative variation is: 92.1%
PC 5 accounts for 4.8% of variation; cummulative variation is: 96.9%
PC 6 accounts for 2% of variation; cummulative variation is: 98.9%
PC 7 accounts for 0.5% of variation; cummulative variation is: 99.4%
PC 8 accounts for 0.4% of variation; cummulative variation is: 99.8%
PC 9 accounts for 0.2% of variation; cummulative variation is: 100%
PC 10 accounts for 0% of variation; cummulative variation is: 100%
```

A scree plot of the variables shows there is no strong “dog leg” hook to the chart, indicating more of a subjective choice when using the scree plot to select the number of principal components (PC) to use. As expected, Principal Component 1 accounts for the largest amount of variation.

Figure 3



The loadings for the first 5 PCs and the name of the corresponding time-use variable are:

Table 3

PCA Un-rotated Loadings

	0	1	2	3	4	use
0	-0.456322	-0.077780	-0.061925	-0.069384	0.112323	professional
1	-0.456312	0.036983	-0.009606	-0.012894	-0.167306	transport
2	0.416242	0.024514	0.345749	-0.146168	-0.005877	housework
3	0.402458	0.134958	0.120137	-0.258610	-0.294994	childcare
4	0.261863	0.521557	-0.022173	0.139745	-0.122016	shopping
5	0.035394	0.563638	-0.270631	0.049177	0.638841	personal
6	0.274992	-0.455181	-0.363171	-0.117594	0.008455	mealtime
7	0.306118	-0.394952	-0.204512	0.199041	0.476168	sleep
8	0.045262	0.144495	-0.770613	-0.281193	-0.346945	tv
9	0.084366	-0.001524	-0.144229	0.867360	-0.319612	leisure

The first 4 of the 10 principal components capture 92% of the variability. PC1 has its' strongest correlations (albeit not terribly strong) with Professional and Transport having a -.456 correlation with both. This indicates the first principal component will increase as Professional and/or Transport decrease. PC1 has a positive correlation with Housework and Childcare which is nearly as strong as the negative correlation with Transport and Professional. This would indicate the principal component will increase when the Housework and Childcare variables increase. So, as a subject spends less time on Professional and Transport, the PC increases; and as they spend more time doing Housework or Childcare the PC increases. This component seems to be related to time spent at home, possibly a "Homemaker" component.

PC2 is most strongly correlated with Shopping and Personal, increasing as they increase. Tempting to think of this as the "Persons of Leisure" component. PC3 has a strong negative correlation (-.771) with TV, indicating an increase in the component with a decrease in TV watching. PC4 has a very high correlation with Leisure, at .867. And finally, PC5 has positive correlations with Personal and Mealtime; perhaps the relaxed, "Time for oneself" component.

Principal Components Analysis — Rotated

Creating a subset of the principal components loadings, which only has the first 2 PCs, I applied the Varimax rotation to it, and got:

Table 4 – Rotated Prinicpal Component Loadings

Varimax loadings PC1 & PC2:

	0	1	use
0	-0.458880	-0.060897	professional
1	-0.454637	0.053787	transport
2	0.416863	0.009146	housework
3	0.407161	0.120023	childcare
4	0.280921	0.511545	shopping
5	0.056157	0.561949	personal
6	0.258017	-0.465014	mealtime
7	0.291344	-0.405973	sleep
8	0.050561	0.142727	tv
9	0.084252	-0.004635	leisure

Varimax PCA variable uniquenesses:	[0.786 0.79 0.826 0.82 0.659 0.681 0.717 0.75 0.977 0.993]
Varimax PCA variable communalities:	[0.214 0.21 0.174 0.18 0.341 0.319 0.283 0.25 0.023 0.007]

My interpretation of the principal axis for the rotated components hasn't changed from the un-rotated interpretation. The main encompasses Professional/Transport – Housework/Childcare; where a decrease in time spent doing Professional/Transport increases the results for the component and an increase in Housework/Childcare increase the component. If I had to label this axis, I think it might be something like "Time spent Working Inside the Home". The second axis again has a negative relationship with Mealtime/Sleep, and a positive correlation with Shopping/Personal. As a study subject sleeps less, or spends less time eating, the PC increases; as it also does when spending more time shopping or engaging in "personal" time. A possible label for this might be "Non-work activity, outside the home". I say, "outside the home", since a reduction in in-home activities increases the PC.

Factor Analysis

Performing Factor Analysis on the data, as per the sample we were given, yields the following values for the loadings for the first 5 Factors

Table 5 – Factor Loadings

FA Loadings

[-0.997	0.069	-0.015	-0.000	0.000]
[-0.949	-0.069	0.189	0.015	-0.009]
[0.893	-0.272	-0.189	-0.293	0.056]
[0.841	-0.391	-0.079	0.048	0.177]
[0.624	-0.378	0.368	0.224	-0.459]
[0.091	-0.261	0.222	0.380	-0.598]
[0.503	0.473	-0.470	0.261	0.447]
[0.613	0.681	-0.400	-0.000	-0.000]
[0.053	-0.059	-0.068	0.977	0.070]
[0.281	0.592	0.756	-0.000	0.000]]

Reducing this to 2 factors, per the directions in item (4) of the assignment, and adding a column to show the variables each loading is associated with, gives us:

Table 6 – Un-rotated Factor loadings and associated variable names

FA Loadings

	0	1	use
0	-0.997483	0.069288	professional
1	-0.949157	-0.068917	transport
2	0.892666	-0.272121	housework
3	0.840976	-0.391049	childcare
4	0.623856	-0.377586	shopping
5	0.090813	-0.261411	personal
6	0.502543	0.472737	mealtime
7	0.613190	0.680991	sleep
8	0.052871	-0.058633	tv
9	0.280891	0.591742	leisure

FA Variable uniquenesses:	[0.000 0.094 0.129 0.140 0.468 0.923 0.524 0.160 0.994 0.571]
FA Variable communalities:	[1.000 0.906 0.871 0.860 0.532 0.077 0.476 0.840 0.006 0.429]

If we rotate these factors, using the Varimax rotation, we get

Table 7 – Factor loadings after applying Varimax rotation

Varimax Rotated FA factor loadings:

	0	1	use
0	-0.943370	-0.331400	professional
1	-0.844263	-0.439168	transport
2	0.927449	0.103641	housework
3	0.927083	-0.026034	childcare
4	0.722382	-0.099656	shopping
5	0.186912	-0.204075	personal
6	0.274242	0.633104	mealtime
7	0.293370	0.868150	sleep
8	0.071768	-0.032901	tv
9	0.023584	0.654601	leisure

FA Rotated Varimax variable uniquenesses:	[0.000 0.094 0.129 0.140 0.468 0.923 0.524 0.160 0.994 0.571]
FA Rotated Varimax variable communalities:	[1.000 0.906 0.871 0.860 0.532 0.077 0.476 0.840 0.006 0.429]

Note: that the rotation, as expected, had no effect on the Uniqueness and Communality measures.

If I were going to “name” the axis associated with these factors, the first seems to me to be the Professional-Transport versus the Homemaker-Caregiver-Shopping. I say versus, in the sense the Professional-Transport values are negatively correlated with the factor, while the Housework-Caregiver-Shopping values are positively correlated. Perhaps it could be thought of as the “Parenting” time expenditure bucket, since spending less time on Profession/Transport raises the factor value, as does spending more time doing Housework-Caregiver-Shopping. The second axis seems to center around non-work activities, namely transport (negative), sleep, meals, and leisure; I’d interpret this as the “Self-maintenance” time bucket.

Next, let's try an alternate rotation, this time using a matrix that places the communalities on the diagonal of the rotation matrix. We know from class and our jump-start solution that *"communality represents the proportion of variable variance that is common to the factor analytic solution".* So, multiplying the identity matrix by the communality vector gives us a new matrix we can use as the rotation matrix. Using the "dot-product" method we multiply the new rotation matrix by the two factors' loadings to get the new rotated loadings:

Table 8

Loadings after rotation via alternate matrix

	0	1	use
0	-0.997255	0.062751	professional
1	-0.948941	-0.062414	transport
2	0.892463	-0.246446	housework
3	0.840785	-0.354153	childcare
4	0.623714	-0.341960	shopping
5	0.090792	-0.236747	personal
6	0.502428	0.428134	mealtime
7	0.613050	0.616739	sleep
8	0.052859	-0.053101	tv
9	0.280827	0.535911	leisure

This rotation does differ from the Varimax version. In the case of Factor 1, we still have negative Professional/Transport as our highest loadings, followed by Housework/Childcare, but now Shopping has dropped a bit, and Sleep and Mealtime have risen to be high enough to be considered as part of the aggregate to name. I find this collection something of a mash-up, and I'm uncertain how I could appropriately label the collection to be meaningful. Since only 3 of the 10 time-buckets are below my personal .40 cut-off, this is not a very useful view of the variables. Factor 2 has Mealtime/Sleep/Leisure as its variables. I suppose one way you could label these would be as "Adult" for factor 1 and "Non-Adult" for factor 2, based on the time spent in the various activities. Non-adults not typically spending time in Professional, Housework, or Childcare activities.

Model Comparison and Recommendation

In an intellectual way, I understand the motivation of Social Science to use factors and factor analysis. There is an appeal to the notion of being able to capture data on underlying, and unobservable traits through manipulation of observed data. However, I find factors, and Factor Analysis even more difficult to explain than Principal Components. I am inclined to stick with the rotated (or even un-rotated) PCA.

Comparing the rotated PC loadings to the original we see there is little difference. The table below shows the arithmetic difference between the original PCA loadings and the once produced by the Varimax rotation:

Table 9 - Original PC Loadings minus Rotated Loadings

Difference between Rotated and Original loadings

	0	1	use
0	-0.002558	0.016883	professional
1	0.001674	0.016804	transport
2	0.000621	-0.015368	housework
3	0.004704	-0.014935	childcare
4	0.019058	-0.010013	shopping
5	0.020764	-0.001689	personal
6	-0.016975	-0.009832	mealtime
7	-0.014775	-0.011021	sleep
8	0.005298	-0.001768	tv
9	-0.000114	-0.003111	leisure

The impact of rotation seems very small, at least for the first 2 principal components. Given that the first principal component is supposed to maximize the variance among the data, and the Varimax rotation is supposed to maximize the sum of the variances of the squared loadings, it is not surprising that the difference between the original principal components and the rotation is small.

In suggesting a solution to management for use in targeting marketing, I believe management is likely to find either the PCA or FA solution confusing. Neither of these techniques have analogy to other fields that would render the technique easy to explain. To some degree, I'm certain the real answer depends on what marketing is trying to do. However, needing to select one, I believe that the rotated principal components solution is slightly easier to explain.

With the PCA solution, there are fewer variables per-component that are above (my arbitrary) .40 threshold. In expressing PC1 as the "Homemaker" or "Mom" axis, it is pretty easy to follow how less time commuting or at work could increase the PC, while more time in housework or childcare also raises the PC. Contrast this with the Verimax FA solution, where we have an F1 which has 5 variables: Professional, Transport, Homemaker, Caregiver and Shopping. It is already harder to explain; the addition of Shopping makes it much harder to provide a good taxonomy for this axis. What sort of shopping? Grocery shopping fits with the "Mom" label, but shopping as recreation does not. F1 is more ambiguous in my mind. Comparing F1 to the alternate-F1 (the common variance matrix rotation) we get 7 variables, and the only interpretation I could come up with is to call that the "Adult" factor.

Conclusion

Both Factor Analysis and Principal Component Analysis are non-intuitive. Using either will require a fair amount of explanation should management desire a deep-dive into the methodology. However, based on this exercise, I am unconvinced of the usefulness of Factor

Analysis. I don't see that the factors added anything more illuminating to the interpretation of the results. I think that if Factor Analysis rotations yield variable loadings which map *cleanly* to the question you are investigating, then by all means, consider using FA for your solution. Personally, I'd stick with the classic solution and use PCA. I didn't feel that the factors provided any extra insight into the data that the component missed. Also, based on this exercise, I'm not certain that PCA rotation buys you anything over the un-rotated principal components; using plain PCA appears to be sufficient to the task.