# Principal Component Analysis and Factor Analysis of the Time-Use Case Dataset

Prepared by Amee Amin

Assignment 7

MSPA Course 410, Summer 2017

Professors Chad R. Bhatti and Tom Miller

## Introduction

Reducing dimensionality is an invaluable method for making large datasets more manageable during data analysis. However, although there are many methods for reducing dimensionality, there is no standard or correct approach. In this analysis, we compare and contrast the methods of principal component analysis and factor analysis. We analyzed a time-use dataset consisting of 10 time-use variables and 28 observations from a time-use survey study conducted across many countries in the mid-1970s. We found that PCA maximized the total variance in our dataset by selecting components that strongly associated with leisure, tv, sleep, professional, personal, and shopping. In comparison, we used Factor Analysis to maximize the shared variance in our dataset by selecting five factors that strongly correlated with leisure, tv, sleep, personal, professional, housework, transport, and childcare (eight of the ten time-use variables). Overall, we recommend that factor analysis be used by management to more easily explain the underlying latent, common variables in our dataset.

## Section 1: Sample Data and Exploratory Data Analysis

Our time-use dataset contained 10 time-use variables with 28 observation groups. Demographic data on group, gender, professional work status, marital status, and country were removed while cleaning the data. The dataset was created for the purpose of explaining and testing classification techniques. The time-use variables describe general categories of activities carried out in daily life: professional (time spent working), transport (time spent in transportation), housework (time spent doing housework), childcare (time spent taking care of children), shopping (time spent shopping), personal (time spent doing personal activities), mealtime (time spent eating), sleep (time spend sleeping), tv (time spent watching tv), and leisure (time spend doing activities for leisure). The time-use variables reflect the average number of minutes spent in a fixed-time observation period of 40 hours. We also noted that exercise, a common daily activity, was not included in this dataset. The average amount of time spent on activities ranged from 0 to approximately 850 minutes.

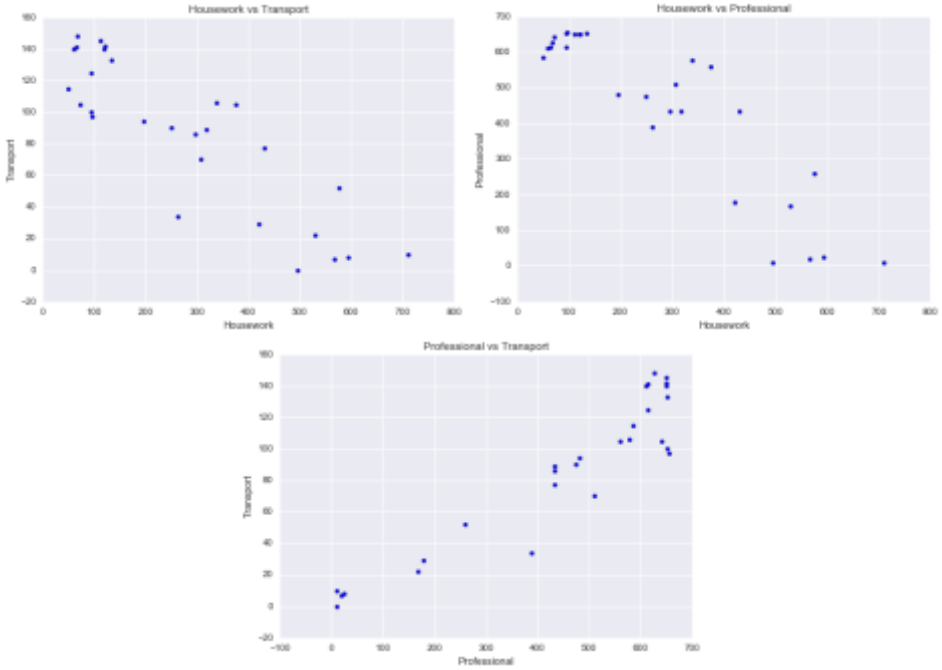
We looked at a correlation matrix of the 10 time-use variables to quickly explore the relationships between variables. Professional and transport appear to be strongly positively correlated, while professional and housework appear to be strongly negatively correlated. Housework and transport are therefore also strongly negatively correlated as well. Childcare expectedly has a strongly negative correlation with professional and transport and a strongly positive correlation with housework. Shopping has a negative correlation with professional and transport, and a positive correlation with housework and childcare. Personal time has a negative correlation with professional, transport, and housework, while it has a positive correlation with childcare and shopping. The definition of personal time should be furthered explored to understand how subjects defined personal time for themselves. Mealtime negatively correlates with professional, transport, shopping, and personal, while it positive correlates with housework and childcare. Sleep negatively correlates with professional, transport, shopping and personal, and positively correlates with housework, childcare, and mealtime. Both sleep and tv have weak correlations with other variables. We explore the strongest correlations further through the scatterplots below.

**Table 1.1: Correlation Matrix of Time-Use Variables**

| | professional | transport | housework | childcare | shopping | personal | mealtime | sleep | tv | leisure |
|---|---|---|---|---|---|---|---|---|---|---|
| **professional** | 1.000000 | 0.939146 | -0.906398 | -0.864779 | -0.654015 | -0.112085 | -0.461478 | -0.558446 | -0.055886 | -0.250538 |
| **transport** | 0.939146 | 1.000000 | -0.870412 | -0.810035 | -0.503057 | -0.077350 | -0.610170 | -0.704684 | -0.041169 | -0.164435 |
| **housework** | -0.906398 | -0.870412 | 1.000000 | 0.861274 | 0.499720 | -0.039976 | 0.358374 | 0.437793 | -0.205751 | -0.053089 |
| **childcare** | -0.864779 | -0.810035 | 0.861274 | 1.000000 | 0.541772 | 0.117891 | 0.364119 | 0.280785 | 0.121612 | -0.054544 |
| **shopping** | -0.654015 | -0.503057 | 0.499720 | 0.541772 | 1.000000 | 0.590629 | -0.182912 | -0.021966 | 0.218571 | 0.229963 |
| **personal** | -0.112085 | -0.077350 | -0.039976 | 0.117891 | 0.590629 | 1.000000 | -0.353476 | -0.211088 | 0.324885 | 0.038271 |
| **mealtime** | -0.461478 | -0.610170 | 0.358374 | 0.364119 | -0.182912 | -0.353476 | 1.000000 | 0.818196 | 0.317928 | 0.065580 |
| **sleep** | -0.558446 | -0.704684 | 0.437793 | 0.280785 | -0.021966 | -0.211088 | 0.818196 | 1.000000 | 0.019653 | 0.272753 |
| **tv** | -0.055886 | -0.041169 | -0.205751 | 0.121612 | 0.218571 | 0.324885 | 0.317928 | 0.019653 | 1.000000 | -0.071656 |
| **leisure** | -0.250538 | -0.164435 | -0.053089 | -0.054544 | 0.229963 | 0.038271 | 0.065580 | 0.272753 | -0.071656 | 1.000000 |

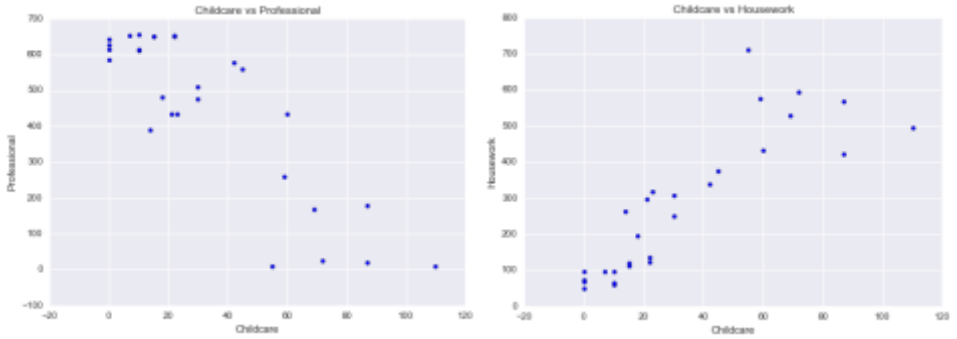
In the scatterplots below, we see a clear negative linear relationship between housework versus transport (left) and housework versus professional (right), and a clear positive linear relationship between professional and transport (center, below). We can infer that as more time is spent away from home doing professional work or in transportation, less time is spent at home doing housework. The range of time spent doing housework or professional work is between 0 and 700 minutes, while the range of time in transportation is between 0 and 160 minutes.

**Fig 1.1: Scatterplots of Housework vs Transport (left), Housework vs Professional (right), and Professional vs Transport (center, below)**



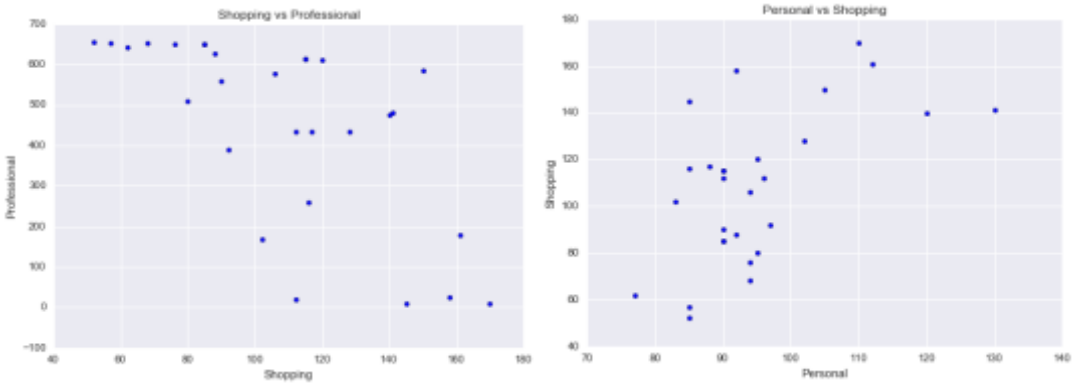
The scatterplots of childcare versus professional and housework (below) show opposite but expected linear correlations. The more time that is spent doing childcare, the less time spent doing professional work and more time spent doing housework. We might infer that this pattern reflects a stereotype of parent roles, where a parent who must spend more time working professionally is unable to spend as much time doing housework.

**Fig 1.2: Scatterplots of Childcare vs Professional (left) and Childcare vs Housework (right)**



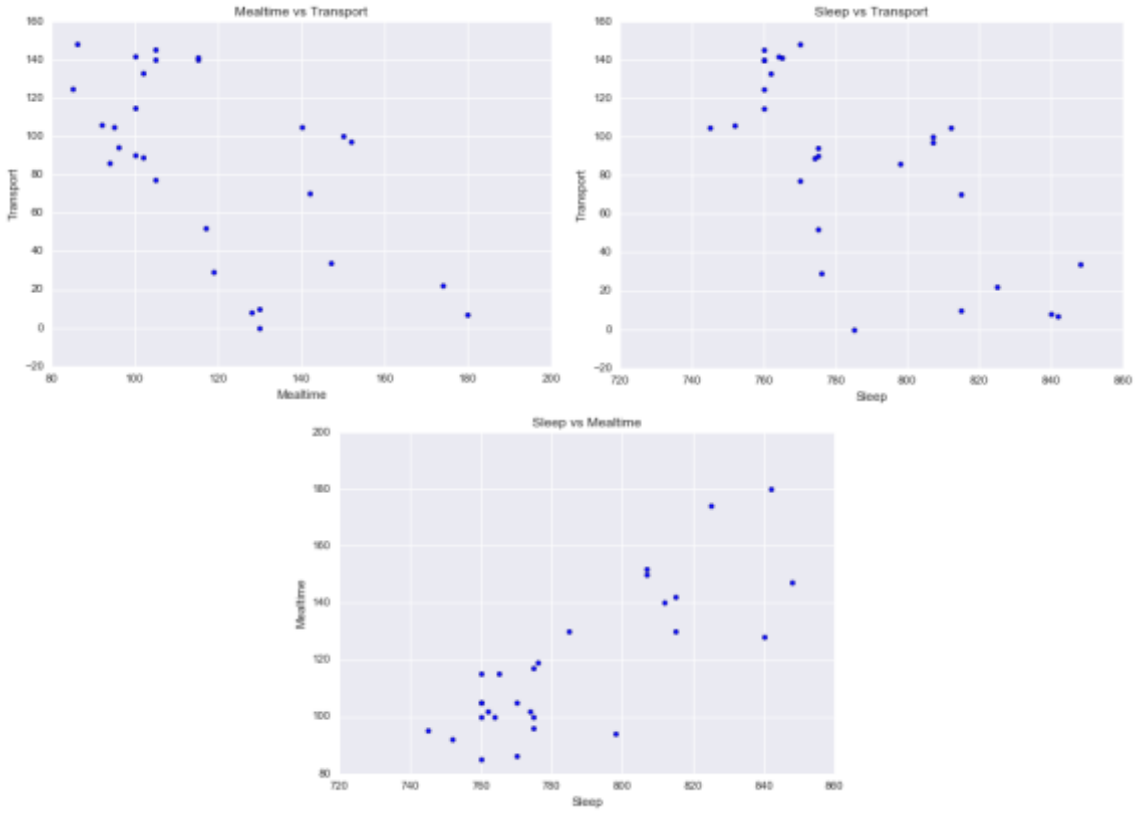

The scatterplot of shopping and professional (left, below) is negatively correlated but not strongly, indicating to us that shopping may still be a priority activity for those who spend a lot of time working professionally. Personal and shopping are positively correlated, but the range of personal time appears to be smaller than that of other variables. Personal time is clustered between 70 and 100 minutes. This indicates that subjects, regardless of their time spent on other activities (such as professional work or housework), tend to have the same amount of personal time. Shopping is also clustered between 40 and 120 minutes.

**Fig 1.3: Scatterplots of Shopping vs Professional (left) and Personal vs Shopping (right)**




Transport shows a negative correlation with both mealtime and sleep in the scatterplots below, while sleep and mealtime have a positive correlation. The range of sleep is between 740 and 860 minutes, which is expected given a sample population of human subjects. We did not have any assumptions about the cause of the relationship between mealtime and transport or mealtime and sleep, and there may be other causal factors that influence trends in mealtime, transport, and sleep.

**Fig 1.4: Scatterplots of Mealtime vs Transport (left), Sleep vs Transport (right), and Sleep vs Mealtime (center, below)**




Overall, there appear to be some strong correlations between the time-use variables and we may expect collinearity in the data. In the next sections, we explore different methods of reducing dimensionality and avoiding collinearity.

## Section 2: Principal Component Analysis – Original/Unrotated

We applied PCA to our explanatory variables using the sklearn package. The principle components are the dimensions along which our data is most differentiated, as expressed by one, two, or more variables. In layman's terms, the principal components reflect an underlying dimension that summarizes or accounts for the original variables. The total explained variance ratio for each component is listed below.

**Table 2.1: PCA Output Explained Variance Ratio**

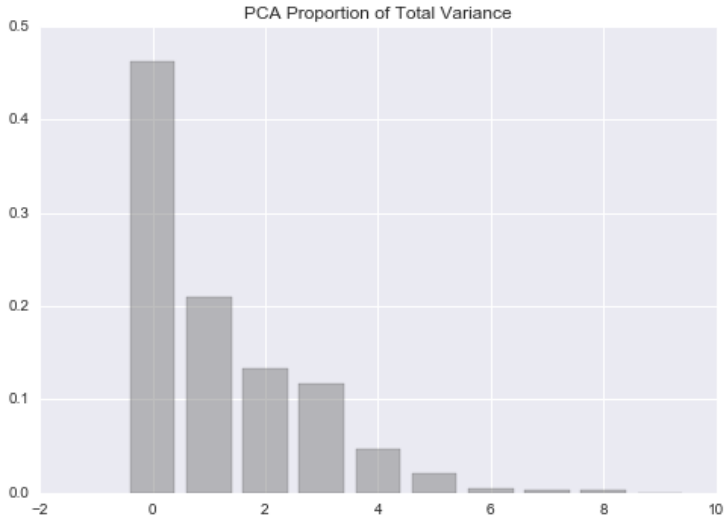
```
[   4.61720629e-01   2.09997139e-01   1.33037184e-01   1.16384699e-01
    4.76070498e-02   2.03274714e-02   4.74636253e-03   3.72923455e-03
    2.24738994e-03   2.02839795e-04]
```

The explained variance ratio represents the proportion of variance explained by each of the selected components. The output above is ordered to reflect each component's explained

variance ratio in decreasing order. The first component explains the most variance in our data, approximately double that of the second component. The tenth component appears to explain a very minute amount of variance in our data.

Looking at a Scree Plot below, there is a huge jump in explained variance ratio from the first component to the second component. The difference between the first component and second component is greater than that of any further sequential components. The third and fourth components appear to explain a similar amount of variance in our data. Nearly all of the variance is captured by the first six principal components.

**Fig 2.1: Scree Plot of PCA**




The cumulative explained variance ratio, shown in the table below, reveals the cumulative proportion of variance explained by each of the selected components. Generally, researchers look to determine the principal components that explain 0.95 of the variance. The first component reflects 0.46 of the variance in our data. The first and second components reflect 0.67 of the variance in our data. The first three components reflect 0.80 of the variance, the first four explain 0.92 of the variance, and the first five explain 0.96 of the variance in our data. As confirmed by the skree plot above, components seven through ten reflect less than .1 of variance in our data.

**Table 2.2: PCA Output Cumulative Explained Variance Ratio**

```
[ 0.46172063  0.67171777  0.80475495  0.92113965  0.9687467   0.98907417
  0.99382054  0.99754977  0.99979716  1.        ]
```

Each principal component consists of a set of factor loadings. The factor loadings for the first five principal components are listed in the table below. A factor loading represents the correlation between the original variable and the factor (the new underlying dimension).

**Table 2.3: PCA Output Factor Loadings**

```
[[-0.456 -0.078 -0.062 -0.069  0.112 -0.054 -0.054 -0.441  0.023  0.751]
 [-0.456  0.037 -0.010 -0.013 -0.167  0.014  0.241  0.493 -0.674  0.079]
 [ 0.416  0.025  0.346 -0.146 -0.006  0.079 -0.180  0.543  0.121  0.580]
 [ 0.402  0.135  0.120 -0.259 -0.295 -0.553  0.418 -0.303 -0.259  0.109]
 [ 0.262  0.522 -0.022  0.140 -0.122  0.604 -0.101 -0.322 -0.368  0.101]]
```

In the next section, we rotate the principal component solutions to make them more interpretable.

## Section 3: Principal Component Analysis – Rotated

Because the principal component solutions are difficult to interpret, we conducted an orthogonal rotation called varimax. The factor axes are rotated while maintaining their orthogonality, so the factor scores remain uncorrelated. The individual factor loadings are moved in the direction of plus or minus one or zero, which shows whether a variable is strongly associated with a principal component factor or not. The rotated factor loadings are listed in the table below.

**Table 3.1: Varimax Factor Loadings**

```
('Varimax factor loadings: ', array([[-0.000, -0.000,  0.000,  0.000, -0.000,  0.000, -0.000, -0.000,
        -0.000,  1.000],
       [ 0.000, -0.000, -0.000, -0.000, -0.000,  0.000,  0.000, -0.000,
         -1.000, -0.000],
       [-0.000,  0.000,  0.000,  0.000,  0.000, -0.000, -0.000,  1.000,
         -0.000,  0.000],
       [ 1.000,  0.000, -0.000,  0.000,  0.000, -0.000,  0.000,  0.000,
          0.000,  0.000],
       [ 0.000,  0.000, -0.000, -0.000, -0.000,  1.000,  0.000,  0.000,
          0.000, -0.000],
       [-0.000, -0.000,  0.000,  0.000,  1.000,  0.000, -0.000, -0.000,
         -0.000,  0.000],
       [ 0.000, -0.000,  0.000, -0.000, -0.000,  0.000, -1.000, -0.000,
         -0.000, -0.000],
       [ 0.000, -1.000, -0.000, -0.000, -0.000,  0.000,  0.000,  0.000,
          0.000, -0.000],
       [-0.000,  0.000, -1.000,  0.000,  0.000, -0.000, -0.000,  0.000,
          0.000,  0.000],
       [-0.000, -0.000,  0.000,  1.000, -0.000,  0.000, -0.000, -0.000,
         -0.000, -0.000]]))
```

The first component, which explained almost half of the variance in our data, is most strongly associated with leisure in the negative direction. This is consistent with our exploratory data analysis that showed leisure had the weakest correlations with other variables, and therefore explains the most variance. The second component, which captured approximately 0.2 of the variance, is tv in the negative direction. The third component, which captured approximately 0.13 of the variance, is most strongly associated with sleep in the positive direction. The fourth component, which captured approximately 0.12 of the variance, is the most strongly associated with professional in the positive direction. The fifth component, which captured approximately 0.04 of the variance, is most strongly associated with personal in the positive direction. The sixth component, which captured approximately 0.02 of the variance, is most strongly associated with shopping in the positive direction.

Given that transport, housework and childcare were most strongly correlated with professional, it is not unexpected that they are not strongly associated with the first six components. Additionally, we observed that mealtime strongly correlated with sleep, and sleep is associated with the third component. The purpose behind PCA is to find the components that maximize the total variance in our dataset, therefore we expected the factor loadings to be most associated with explanatory variables that are not strongly correlated with each other.

## Section 4: Factor Analysis

In this section we explore using factor analysis to similarly reduce dimensionality in our dataset. Factor analysis differs in that its goal is to determine the "common underlying factors" that maximize shared variance, instead of total variance. In factor analysis, you assume that the latent constructs are the actual causal factors underlying the collinearity among your variables. Therefore, factor analysis looks not only at a covariance matrix of variables but also a covariance matrix of noise.

We performed a factor analysis to find 5 underlying factors that maximize shared variance. The proportion of variance explained by each of the rotated factors are listed below. The first factor explains almost double that of the second factor. The second factor explains almost double that of the third, fourth, or fifth factors. The third, fourth, and fifth factors explain a similar amount of variance.

**Table 4.1: Factor Analysis Explained Variance**

| | | | | |
|---|---|---|---|---|
| [ 3.811 | 2.007 | 1.109 | 1.220 | 1.181] |

The table below shows the rotated factor loadings which represent both how the variables are weighted for each factor and also the correlation between the variables and the factor. Factor one has a strong negative correlation with professional and transport, and a strong positive correlation with housework and childcare. The relationships between variables are similar to those observed during exploratory data analysis. Factor two has a strong positive correlation with sleep. Factor three has a strong positive correlation with leisure. Factor four has a strong positive correlation with tv. Lastly, factor five has a negative correlation with personal. A clear difference in the factor analysis output is that we see how professional, housework, childcare, and transport are reflected in one factor because of their shared variance, as opposed to PCA, where we only see professional strongly associated with the components in order to maximize total variance.

**Table 4.2: Factor Analysis Loadings**

```
('Varimax factor loadings: ', array([[-0.917, -0.313, -0.191, -0.041,  0.150],
       [-0.818, -0.508, -0.084, -0.014,  0.089],
       [ 0.945,  0.209, -0.106, -0.215, -0.020],
       [ 0.932,  0.058, -0.070,  0.152,  0.005],
       [ 0.618, -0.189,  0.211,  0.172, -0.660],
       [ 0.055, -0.197,  0.028,  0.263, -0.718],
       [ 0.290,  0.776,  0.006,  0.325,  0.417],
       [ 0.261,  0.951,  0.153, -0.043,  0.043],
       [-0.025,  0.086, -0.048,  0.970, -0.138],
       [ 0.016,  0.124,  0.989, -0.044, -0.065]]))
```

Because we used an orthogonal rotation in the factor analysis, the same number should appear along the diagonal in the factor correlation matrix (below). This is called a diagonal matrix. Our factor correlation matrix shows a “1” along the diagonal and the rest of the elements are “0”. This confirms that our factors are uncorrelated.

**Table 4.3: Factor Correlation Matrix**

```
[ 1.000 -0.000  0.000  0.000 -0.000]
[-0.000  1.000 -0.000 -0.000 -0.000]
[ 0.000 -0.000  1.000 -0.000  0.000]
[ 0.000 -0.000 -0.000  1.000 -0.000]
[-0.000 -0.000  0.000 -0.000  1.000]]
```

The table below displays the communalities for each variable, or the proportion of each variable’s variance that can be explained by the “underlying” latent factors. Uniqueness is 1-communality. The variance for professional, sleep, and leisure is fully explained by the factors. The variance for transport, housework, childcare, shopping, mealtime, and tv is almost fully explained by the factors (at least 0.90 of their variance). Last, only 0.628 of the variance for personal is explained by the factors. Overall, the factors capture a majority of the variance explained by the time-use variables.

**Table 4.4: Varimax Variable Uniqueness / Communalities**

```
('Varimax variable uniquenesses: ', array([ 0.000,  0.058,  0.005,  0.100,  0.072,  0.372,  0.035,  0.000,
       0.031,  0.000]))
('Varimax variable communalities: ', array([ 1.000,  0.942,  0.995,  0.900,  0.928,  0.628,  0.965,  1.000,
       0.969,  1.000]))
```

## Section 5: Model Comparison and Recommendation

We recommend that the Factor Analysis rotated solution be used by management because it is easier to interpret how each factor represents a different latent relationship among initial explanatory variables.

The strong relationship between professional, transport, housework, and childcare is clearly represented by the first factor in our Factor Analysis. Each factor loading similarly shows not only the relationship between variables and the factor but the relationship between variables (Table 4.2). On the other hand, our Principal Component Analysis seeks to maximize total variance and therefore only strongly reflects one of the four variables. Conceptually, after

observing the relationship between variables in an exploratory data analysis, we think that observing how the variables are, in some sense, “combined” into underlying common variables is easier with Factor Analysis.

We would also recommend that the category of personal is further explored, either by refining the definition for future surveys or better understanding what subject groups detailed as “personal” activities.

## Conclusion

In conclusion, principal component analysis and factor analysis are both invaluable methods for reducing dimensionality in our dataset by uncovering the hidden relationships between explanatory variables. The choice to maximize total variance or shared variance depends on how well one needs to explain the relationships among variables, in their relationship to the principal components or factors. The time-use dataset shows that there are patterns in the time spent on general activities, such as the time spent on work and the time spent in transportation, and dimensionality-reduction methods can help elucidate the underlying relationships among activities.