Derived Attributes and Dimensionality Reduction

T. Evgeniou

What is this for?

One of the key steps in Data Analytics is to generate meaningful attributes starting from possibly a large number of **raw attributes**. Consider for example the case of customer data, where for each customer we have a very large number of raw attributes (which could be, for example, *independent variables* in the case of a regression used to predict acquisition, purchases, or churn) ranging from demographic information, to what products they bought in the past, who their friends on various social media sites are, what websites they visit more often, how they rated a number of movies or products in general, whether they use their mobile phone mostly the weekends or in the mornings, how they responded to various surveys, etc. One can easily end up with tens if not thousands of such raw attributes, for thousands or millions of customers. In such cases it is virtually impossible to use some of the "advanced" methodologies developed for data analytics, or even simple ones such as linear regression. This is not only because of compuational reasons but, more important, because of statistical/mathematical reasons as, doing so entails a high risk that the estimated models (e.g. consumer behavior models) may be of very low quality in a statistical hence also practical sense. Moreover, the insights developed may not be practical or actionable as they may correspond to complicated statements involving a very large number of raw customer attributes.

In such situations, which arise almost always in practice, one needs to spend a lot of creative effort and time - based on a deep contextual knowledge - to generate new attributes (e.g. "this customer is price sensitive", or "this customer likes comfort", or "this customer is status conscious", etc) from the original raw ones, which we call here **derived attributes**. These attributes can be generated manually, or using what we call **data or dimensionality reduction statistical techniques**.

We will consider a specific family of such statistical techniques which we will broadly call **Factor Analysis techniques**. Such techniques are often used at the early stages of a data analytics project, for example before running a regression with many independent variables (the raw attributes in that case), in order to summarize information (the variation) in correlated raw attributes using a smaller number of manageable **factors** - which are typically uncorrelated or independent. In the process one decreases the number of raw attributes while **keeping most of the information in the data, in a statistical sense**.

Such derived variables are usually useful also for managerial interpretation and action. For example, when analyzing survey data from students regarding the quality of a school, the quality of education may be a useful variable, but in a student survey one could instead ask several questions related to this: (1) Breadth of courses; (2) Depth of courses; (3) Quality of Instruction; (4) Practicality of coursework, etc. A "good linear combination" of these raw attributes may be more managerially useful in understanding student perceptions of the quality of education than each of these variables alone. If indeed these variables are highly related to the underlying construct of "quality of education", Factor analysis will be able to summarize the information into such a single factor, while capturing most of the information in those "raw" attributes.

> Before proceeding on understanding the statistical techniques considered here, it is important to note that this is not the only approach to generating meaningful derived attributes from large numbers of raw ones: there is always "the manual approach"" based on contextual knowledge and intuition, which can probably take a data analytics project already very far. However, in terms of mathematical techniques used in data analytics, factor analysis is one of the key ones when it comes to generating new meaningful derived attributes from the original raw ones.

An Example

There are many dimensionality reduction statistical methods. In this note we will go through the steps of one specific approach. We will do so using a simple example. For this example the "meaningful derived variables" will seem to be straightforward to design, which can help us with the intuition of what the method does. Having said this, before reading the analysis below, do try to think what "derived attributes" one could get from the raw ones below. You will see that even in this case it is not as obvious and you will most likely disagree with your colleagues about what the derived attributes should be: so we will let the numbers and statistics help us be more *objective and statistically correct*.

The "Business Decision"

We consider the core decision of an MBA admissions committee: *which applicants should we accept in the MBA program?* The school is interested in predicting the MBA participant's success in the future before offering admission.

The Data

To make this decision, the committee uses a number of data about the applicants. Let us consider for example the following attributes in evaluating an MBA application (of course in practice many more can be considered):

1. GPA
2. GMAT score
3. Scholarships, fellowships won
4. Evidence of Communications skills (debating competition, personal interview score)
5. Prior Job Experience
6. Organizational Experience
7. Other extra curricular achievements

Let us assume that this data is converted into a numerical scale from 1-7. For example: a numerical rating may be given to the fellowships based on the prestige and the number of fellowships won. Job experience may be rated on the number of years on the job, with a numerical weighting for the level of the job in the managerial ladder.

This is how the first 10 data looks:

	GPA	GMAT	Fellow	Comm	Job.Ex	Organze	Extra
observation 01	3.0	580	2.0	3.5	5	3.8	4.0
observation 02	3.2	570	2.0	3.8	6	3.8	3.8
observation 03	3.7	690	3.0	3.3	3	3.2	3.6
observation 04	3.9	760	3.0	3.8	5	3.9	3.2
observation 05	2.8	480	2.0	3.2	6	3.8	3.8
observation 06	3.4	520	2.5	2.6	2	2.5	2.4
observation 07	3.6	670	3.0	3.7	4	3.5	2.9
observation 08	3.6	760	3.0	3.9	5	3.3	3.2
observation 09	2.8	380	1.0	2.0	3	2.9	3.1
observation 10	3.6	560	3.0	2.8	2	1.0	2.8

We will see some descriptive statistics of the data later, when we get into statistical analysis.

The Approach

How can this data inform the school's admission decisions? Does this data capture some "derived attributes" that may have a meaning? If you were to derive 2 or 3 attrributes from the data above (by combining them in various ways, for example), what would those be? Which raw attributes would be "linked" with the derived ones you think of? **Try it** intuitively before reading any further...

Intuitively it may seem that the data above capture two fundamental abilities that affect the success of students in their management careers:

1. Basic intelligence
2. Team and Leadership skills.

The school may be interested for example in picking students who score high on these two areas. In this case, of course the admissions commitee in theory could just ask the applicants two questions:

1. "How intelligent are you?"
2. "How strong are your team and leadership skills?"

As you can imagine, asking these questions would not only make the admissions interviewers look naive, but would also lead to very noisy and misleading answers: of course everyone will just answer both questions with the highest mark. So instead of asking these "naive" questions, the school is using *raw attributes/data* like the ones above, which can also be gathered easier. The idea then is to see how this data can be "translated" in meaningful derived attributes that, for example, could capture the "equivalent answers" one could get if one were to ask directly the two naive questions above - or possibly other such "complex"" questions.

> Factor analysis is a statistical approach for finding a few "hidden" derived attributes in data by combining together groups of the original raw attributes in such a way that the least information in the original data is lost - in a statistical sense. It is part of a general class of statistical methodologies used to do what is often called "dimensionality reduction".

Back to our example, if there is some way in which we could reduce the 7 attributes into a smaller set of, say, 2 or 3 attributes, then we can reduce the data to a more understandable form so that our decision making process can be made potentially simpler, more actionable, and easier to interpret and justify - without losing much information in the original data. It is much easier to make tradeoffs between two or three attributes than it is between 10 or 20 attributes (look at any survey or application form and you will see that there are easily more than 20 questions). Hence,

> Data reduction is a very useful step in helping us interpret the data and make decisions.

Like for our example, theory may suggest that there are really one or two basic factors (like intelligence and leadership skills) that lead to success in a management career. The various attributes are really different manifestations of these basic factors. But maybe there are other hidden derived variables (factors) in the data we have: instead of us manually combining raw attributes into meaningful derived ones, which not only is diffucult with many data but also dangerous as we impose our biases, let's get *factor analysis* to do the job for us - and use our intuition and judgment in the process.

Let's now see **a** process for using factor analysis in order to create derived attributes, the goal of this report.

A 6-steps Process for Dimensionality Reduction

> It is important to remember that Data Analytics Projects require a delicate balance between experimentation, intuition, but also following (once a while) a process to avoid getting fooled by randomness and "finding results and patterns" that are mainly driven by our own biases and not by the facts/data themselves.

There is *not one* process for factor analysis. However, we have to start somewhere, so we will use the following process:

1. Confirm the data is metric
2. Decide whether to scale or standardize the data
3. Check the correlation matrix to see if Factor Analysis makes sense
4. Develop a scree plot and decide on the number of factors to be derived
5. Interpret the factors (consider factor rotations - technical but useful)
6. Save factor scores for subsequent analyses

Let's follow these steps.

Step 1: Confirm data is metric

Steps 1-3 are about specific descriptive characteristics of the data. In particular, the methods we consider in this note require that the data are *metric* (step 1): this means not only that all data are numbers, but also that the numbers have an actual numerical meaning, that is 1 is less than 2 which is less than 3 etc. If we have other types of data (e.g. gender, categories that are not comparable, etc), there are other methods to use. However, for now we will only consider a specific method, which we will also mis-use for non-numeric data for simplicity.

The data we use here have the following descriptive statistics:

	min	25 percent	median	mean	75 percent	max	std
GPA	2.5	2.80	3.45	3.31	3.62	3.9	0.47
GMAT	380.0	480.00	575.00	583.50	682.50	760.0	119.44
Fellow	1.0	2.00	2.80	2.45	3.00	3.8	0.91
Comm	2.0	3.18	3.40	3.34	3.73	3.9	0.49
Job.Ex	2.0	3.00	5.00	4.25	5.25	6.0	1.52
Organze	1.0	3.05	3.40	3.20	3.80	3.9	0.73
Extra	2.4	2.88	3.40	3.30	3.80	4.0	0.52

> Note that one should spend a lot of time getting a feeling of the data based on simple summary statistics and visualizations: good data analytics require that we understand our data very well.

Step 2: Scale the data

This step is optional (many methods standardize the data automatically anyway).

Note that for this data, while 6 of the "survey" data are on a similar scale, namely 1-7, there is one variable that is about 2 orders of magnitude larger: the GMAT variable. Having some variables with a very different

range/scale can often create problems: **most of the "results" may be driven by a few large values**, more so that we would like. To avoid such issues, one has to consider whether or not to **standardize the data** by making some of the initial raw attributes have, for example, mean 0 and standard deviation 1 (e.g. `scaledGMAT = (GMAT-mean(GMAT)) / sd(GMAT)`), or scaling them between 0 and 1 (e.g. `scaledGMAT = (GMAT-min(GMAT)) / (max(GMAT)-min(GMAT))`). Here is for example the R code for the first approach, if we want to standardize all attributes:

```
ProjectDatafactor_scaled = apply(ProjectDataFactor, 2, function(r) {
    if (sd(r) != 0)
        res = (r - mean(r))/sd(r) else res = 0 * r
    res
})
```

Notice now the summary statistics of the scaled dataset:

	min	25 percent	median	mean	75 percent	max	std
GPA	-1.72	-1.08	0.31	0	0.68	1.27	1
GMAT	-1.70	-0.87	-0.07	0	0.83	1.48	1
Fellow	-1.60	-0.50	0.39	0	0.61	1.49	1
Comm	-2.73	-0.33	0.13	0	0.80	1.16	1
Job.Ex	-1.48	-0.82	0.49	0	0.66	1.15	1
Organze	-2.99	-0.20	0.27	0	0.82	0.95	1
Extra	-1.75	-0.83	0.19	0	0.97	1.36	1

As expected all variables have mean 0 and standard deviation 1.

While this is typically a necessary step, one has to always do it with care: some times you may want your analytics findings to be driven mainly by a few attributes that take large values; other times having attributes with different scales may imply something about those attributes. For example, when students rate their schools on various factors on a 1-7 scale, if the variability is minimal on a certain variable (e.g. satisfaction about the IT infrastructure of the school) but very high on another one (e.g. satisfaction with job placement), then standardization will reduce the real big differences in placement satisfaction and magnify the small differences in IT infrastructure satisfaction. In many such cases one may choose to skip step 2 for some of the raw attributes. Hence standardization is not a necessary data transformation step, and you should use it judiciously.

Step 3: Check correlations

The type of dimensionality reduction methods we will use here "groups together raw attributes that are highly correlated". Other methods (there are many!) use different criteria to create derived variables. For this to be feasible, it is necessary that the original raw attributes do have large enough correlations (e.g. more than 0.5 in absolute value, or simply statistically significant). It is therefore useful to see the correlation matrix of the original attributes - something that one should anyway always do in order to develop a better understanding of the data.

This is the correlation matrix of the 7 original variable we use for factor analysis (Note: this would be the same for the standardized ones if the standardization is done as above; there is a mathematical reason for this that we will not explore - you could confirm it yourself):

	GPA	GMAT	Fellow	Comm	Job.Ex	Organze	Extra
GPA	1.00	0.90	0.92	0.56	0.15	-0.03	0.01
GMAT	0.90	1.00	0.86	0.78	0.33	0.19	0.16
Fellow	0.92	0.86	1.00	0.59	0.18	0.01	0.02
Comm	0.56	0.78	0.59	1.00	0.60	0.47	0.39
Job.Ex	0.15	0.33	0.18	0.60	1.00	0.80	0.77

	GPA	GMAT	Fellow	Comm	Job.Ex	Organze	Extra
Organze	-0.03	0.19	0.01	0.47	0.80	1.00	0.61
Extra	0.01	0.16	0.02	0.39	0.77	0.61	1.00

There are quite a few large (in absolute value) correlations. For example GPA, GMAT and Fellowship seem to be highly positively correlated - as expected? Maybe those can be grouped in one "factor"? How about "Communication Skills"? Should that also be part of that same factor? With what weights should we combine these raw attributes in groups? Remember, this is a very simple example where one could possibly derive attributes manually. In practice most of the time data are not as easy to understand, with many more than 7 raw attributes. However, even in this simple example people often disagree about how to group the 7 raw attributes!

Let's now see what factor analysis suggests as factors.

Step 4: Choose number of factors

There are many statistical methods to generate derived variables from raw data. One of the most standard ones is **Principal Component Analysis**. This method finds factors, called **Principal Components**, which are **linear combinations of the original raw attributes** so that most of the information in the data, measured using **variance explained** (roughly "how much of the variability in the data is captured by the selected components") is captured by only a few factors. The components are developed typically so that they are **uncorrelated**, leading to *at most as many factors as the number of the original raw attributes, but* so that only a few are needed (the *principal components*) to keep most of the information (variance/variability) in the raw data. For example, for our data we have 7 raw attributes hence we can only have a total of 7 factors/components, each of them being a linear combination of the 7 original raw data.

While there are as many (and for other methods, more) factors as the number of the original raw attributes, since our goal is to have a small(er) number of derived variables/factors, one question is whether we could use only a few of the components without losing much information. When this is feasible, we can say that the original raw attributes can be "compressed" to a few principal components/factors/derived variables. Note that this is not necessarily feasible - e.g. when the original raw attributes are uncorrelated and each one provides "trully different information from all the others" (in this case "different" means "uncorrelated", but other statistical measures of "different" can be used, such as "statistically independent information", leading to other well known dimensionality reduction methods such as *Independent Component Analysis (ICA)*, etc).

When using PCA, we have two measures of "how much of the information (variance in this case) in the original raw data is captured by any of the factors":

a) the *percentage of variance explained*,
b) the *eigenvalue coresponding to the component*.

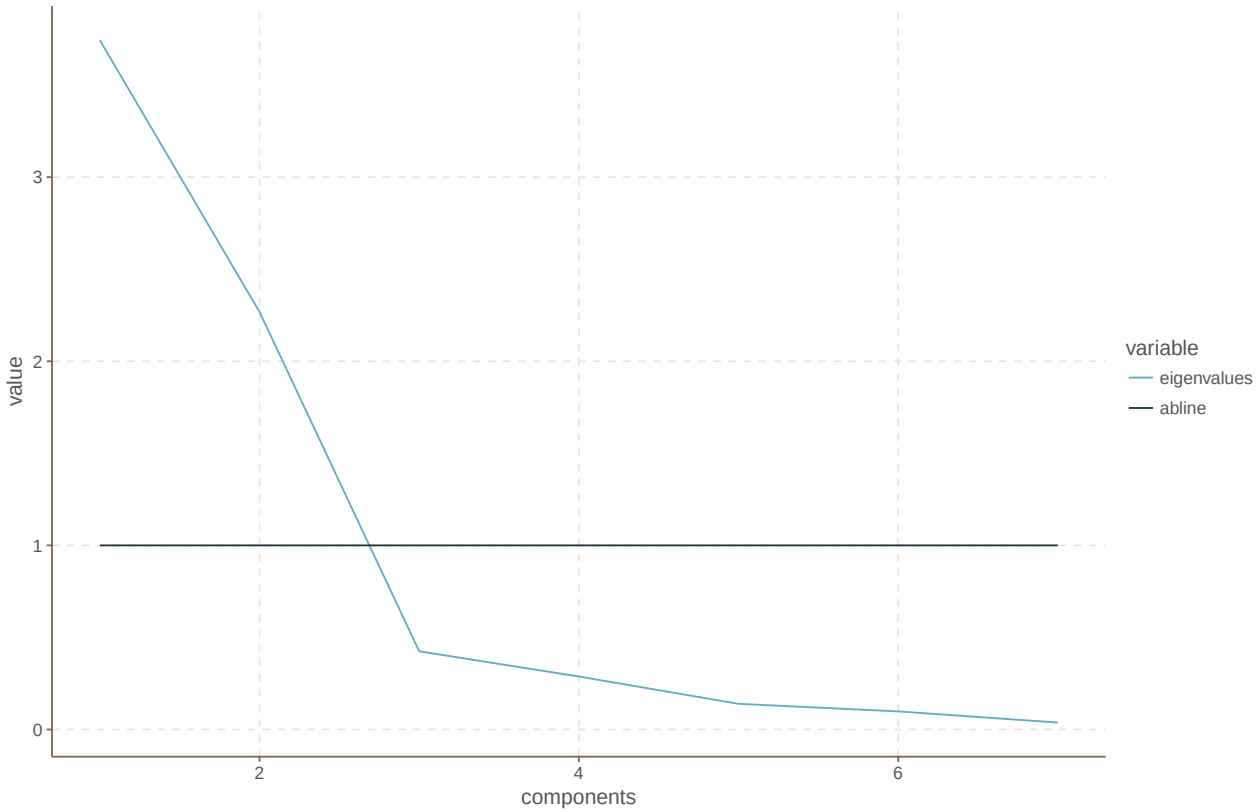
Each factor has an eigenvalue as well as the percentage of the variance explained. The sum of the eigenvalues of the components is equal to the number of original raw attributes used for factor analysis, while the sum of the percentages of the variance explained across all components is 100%. For example, for our data these are:

	Eigenvalue	Pct of explained variance	Cumulative pct of explained variance
Component 1	3.74	53.48	53.48
Component 2	2.27	32.40	85.88
Component 3	0.42	6.07	91.95
Component 4	0.29	4.11	96.06
Component 5	0.14	1.99	98.05
Component 6	0.10	1.41	99.46
Component 7	0.04	0.54	100.00

Note that the "first principal component" has the highest eigenvalue and captures most of the information (variance in this case) of the original raw data. As a rule of thumb, every component with eigenvalue more than 1 "has more information that the average original raw attribute". Typically one uses only these factors, although what is also important to consider is "what total percentage of the variance in the original data is kept when one replaces the original data/attributes with the selected factors/components".

> Two Statistical criteria to select the number of factors/derived variables when using PCA are: a) select components with corresponding eigenvalue larger than 1; b) Select the components with the highest eigenvalues "up to the component" for which the cumulative total variance explained is relatively large (e.g. more than 50%).

One can also plot the eigenvalues of the generated factors in decreasing order: this plot is called the **scree plot**. For our data this plot looks as follows:



> A third rule of thumb to decide how many components to use is to consider only the factors up to the "elbow" of the scree plot.

Based on the three criteria (eigenvalue > 1, cumulative percentage of variance explained, and the elbow of the scree plot), and using our current selection criterion, namely eigenvalue, for this data we can decide to use 2 components only. In practice one may try different numbers of factors/components as one needs to consider not only the statistical rules discussed here, but also the interpretation and actionability of the selected components: as always, data analytics is about both science and art. We consider interpretability of the derived attributes/factors next.

Step 5: Interpret the factors

In practice one would like to have derived variables that use only a few of the original raw attributes - while each new derived variable using different subsets of the original raw attributes. Unfortunately this is not necessarily always the case. However, there are mathematical methods, called "factor **rotations**", which transform the estimated factors into new ones which capture exactly the same information from the raw data but use only few non-overlapping raw attributes. One such rotation often used in practice is called the varimax rotation - but others are also available.

For our data, the 2 selected factors look as follows after the varimax rotation:

	Component 1	Component 2
GPA	0.96	-0.05
GMAT	0.95	0.19
Fellow	0.95	-0.01
Comm	0.70	0.54
Job.Ex	0.19	0.93
Organze	0.01	0.89
Extra	0.01	0.86

To better visualize and interpret the factors we often "supress" loadings with small values, e.g. with absolute values smaller than 0.5. In this case our factors look as follows after suppressing the small numbers:

	Component 1	Component 2
GPA	0.96	
GMAT	0.95	
Fellow	0.95	
Comm	0.70	0.54
Job.Ex		0.93
Organze		0.89
Extra		0.86

Notice that after rotation each factor combines (we say "loads on") only a few of the original raw attributes, making interpretation easier. For example, if we only select the factors with eigenvalue more than 1, in this case we would select 2 factors.

How would you interpret the selected factors?

What Factor Loads "Look Good"? We often use three **Factor Technical Quality Criteria**:

1. For each factor (column) only a few loadings are large (in absolute value)
2. For each raw attribute (row) only a few loadings are large (in absolute value)
3. Any pair of factors (columns) should have different "patterns" of loading

Step 6: Save factor scores

Once we decided the factors to use (for now), we typically replace the original data with a new dataset where each observation (row) is now described not using the original raw attributes but using instead the selected factors/derived attributes. After all this was the goal of this analysis.

The way to represent our observations using the found derived attributes (factors/components) is to estimate for each observation (row) how it "scores" for each of the selected factors. These numbers are called **factor scores**.

Effectively they are the "scores" the observation would take on the factor had we measured that factor directly instead of measuring the original raw attributes.

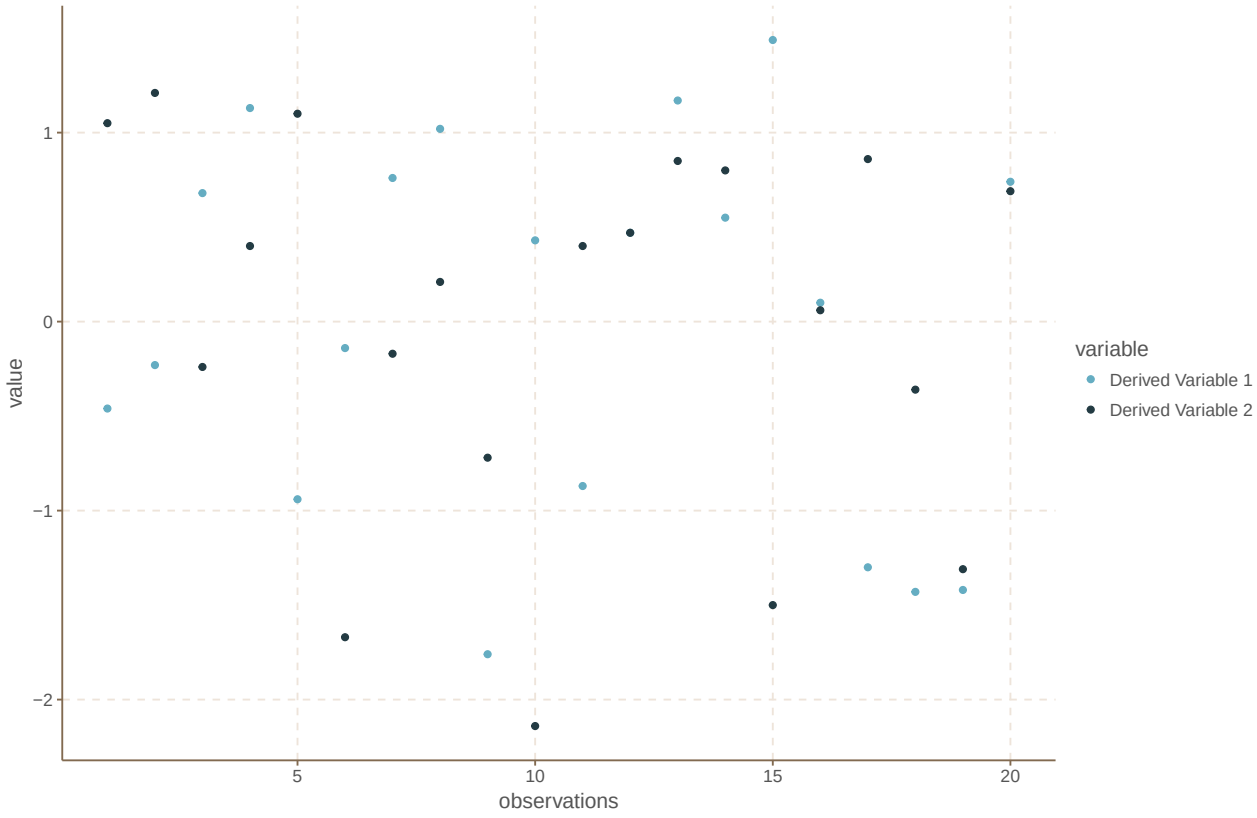
(**Note:** Sometimes for simplicity we represent each selected factor using one of the original raw attributes, typically the one on which the factor has the highest loading on. Although this is not statistically as accurate, it may help with the interpretation of subsequent analyses.)

For our data, using the rotated factors we selected, we can create a new dataset where our observations are as follows (for the first 10 observations):

	Derived Variable (Factor) 1	Derived Variable (Factor) 2
observation 01	-0.46	1.05
observation 02	-0.23	1.21
observation 03	0.68	-0.24
observation 04	1.13	0.40
observation 05	-0.94	1.10
observation 06	-0.14	-1.67
observation 07	0.76	-0.17
observation 08	1.02	0.21
observation 09	-1.76	-0.72
observation 10	0.43	-2.14

Can you describe the observations using the new derived variables? How does each person perform for each of the selected factors?

We now can replace our original data with the new ones and continue our analysis. For example, we can now visualize our original data using only the newly derived attributes:



Remember that we still see a lot of the information in the original data (the total variance explained using 2 factors)

using only this 2-dimensional plot! This is of course only the begining of the analysis using the new attributes. Later on one may need to come back to these tools to generate new derived variables. As always remember that

> Data Analytics is an iterative process, therefore we may need to return to our original raw data at any point and select new raw attributes as well as new factors and derived variables.

Till then...