In **factor analysis** we represent the variables y1, y2, . . . , yp as linear combinations of a few random variables f1, f2, . . . , fm (m < p) called factors. The factors are underlying constructs or latent variables that "generate" the y's. Like the original variables, the factors vary from individual to individual; but unlike the variables, the factors cannot be measured or observed. The existence of these hypothetical variables is therefore open to question.

The two types: exploratory and confirmatory.

- Exploratory factor analysis is if you don't have any idea about what structure your data is or how many dimensions are in a set of variables.
- Confirmatory Factor Analysis is used for verification as long as you have a specific idea about what structure your data is or how many dimensions are in a set of variables.

In this dataset for NYC felony we are trying to observe how different factors are playing an important role in deciding the relationship between the features (in this case, different kinds of crimes) and which ones can be separated out as part of the analysis to reduce the dimensions.

Before we can decide on whether to use priors= SMC we must check the inverse co-relation matrix and if it is invertible, we can decide. From the output below we can confirm it is indeed invertible.

The FACTOR Procedure
Initial Factor Method: Principal Factors

Inverse Correlation Matrix

	Murder	Rape	Robbery	Assault	Burglary	Larceny	LarcenyOfAutomotives
Murder	2.09519	0.19092	-0.07150	-1.38123	-0.07711	0.50967	-0.18793
Rape	0.19092	4.43513	-0.43302	-2.17593	-0.19151	-0.18139	-1.60801
Robbery	-0.07150	-0.43302	8.94126	-7.01300	-1.58467	-0.69765	0.32415
Assault	-1.38123	-2.17593	-7.01300	9.95969	0.45965	0.71611	-0.12942
Burglary	-0.07711	-0.19151	-1.58467	0.45965	3.19498	-1.04416	-1.08599
Larceny	0.50967	-0.18139	-0.69765	0.71611	-1.04416	1.66479	0.40615
LarcenyOfAutomotives	-0.18793	-1.60801	0.32415	-0.12942	-1.08599	0.40615	2.87937

Partial Correlations Controlling all other Variables

	Murder	Rape	Robbery	Assault	Burglary	Larceny	LarcenyOfAutomotives
Murder	1.00000	-0.06263	0.01652	0.30237	0.02980	-0.27290	0.07651
Rape	-0.06263	1.00000	0.06876	0.32739	0.05088	0.06675	0.44997
Robbery	0.01652	0.06876	1.00000	0.74316	0.29649	0.18083	-0.06388
Assault	0.30237	0.32739	0.74316	1.00000	-0.08148	-0.17587	0.02417
Burglary	0.02980	0.05088	0.29649	-0.08148	1.00000	0.45275	0.35805
Larceny	-0.27290	0.06675	0.18083	-0.17587	0.45275	1.00000	-0.18550
LarcenyOfAutomotives	0.07651	0.44997	-0.06388	0.02417	0.35805	-0.18550	1.00000

Communalities (also known as h2) are the estimates of the variance of the factors, as opposed to the variance of the variable which includes measurement error. Initially the communality estimates are set equal to the R^2 between each variable and all others.

The table below shows the initial estimates of communalities:

Prior Communality Estimates: SMC

Murder	Rape	Robbery	Assault	Burglary	Larceny	LarcenyOfAutomotives
0.52271654	0.77452727	0.88815893	0.89959527	0.68700881	0.39932298	0.65270237

Factor loadings: Commonality is the square of the standardized outer loading of an item. Analogous to Pearson's r-squared, the squared factor loading is the percent of variance in that indicator variable explained by the factor. To get the percent of variance in all the variables accounted for by each factor, add the sum of the squared factor loadings for that factor (column) and divide by the number of variables. This is the same as dividing the factor's eigen value by the number of variables.

Interpreting factor loadings: By one rule of thumb in confirmatory factor analysis, loadings should be .7 or higher to confirm that independent variables identified a priori are represented by a particular factor, on the rationale that the .7 level corresponds to about half of the variance in the indicator being explained by the factor. However, the .7 standard is a high one and real-life data may well not meet this criterion, which is why some researchers, particularly for exploratory purposes, will use a lower level such as .6 for the central factor and .25 for other factors. In any event, factor loadings must be interpreted in the light of theory, not by arbitrary cutoff levels.

The table below shows the final estimates of loadings:

Rotated Factor Pattern		
	Factor1	Factor2
Murder	0.71667	-0.21302
Rape	0.85445	0.23022
Robbery	0.89387	0.27044
Assault	0.94941	0.07350
Burglary	0.63445	0.58835
Larceny	0.00585	0.67564
LarcenyOfAutomotives	0.75833	0.19833

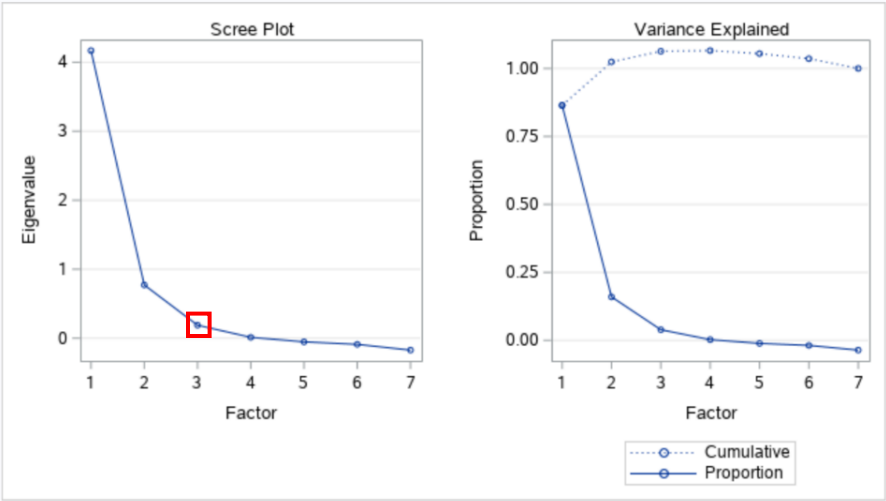
The factors that affect the question the most (and therefore have the highest factor loadings) are bolded. Factor loadings are similar to correlation coefficients in that they can vary from -1 to 1. The closer factors are to -1 or 1, the more they affect the variable. A factor loading of zero would indicate no effect.

For factor 1 all the crimes except ‘Larceny’ are strongly associated. Whereas for ‘Burglary’ it is a pretty close call between Factor 1 and Factor 2 but it doesn’t have a loading of 0.7 or higher for either of those factors and hence, we decide to go with Factor 2 as SAS by default uses the range (0.3 to 0.6) to differentiate factors in cases of conflict. As mentioned in the interpretation earlier, in real life situations a strong association with factors with a value of 0.7 or higher is not always expected in that case we lower the benchmark to 0.6, or better put, the one which is closest to the 0.7 gets the precedence.

The table below shows the proportion of variance explained by each factor:

Eigenvalues of the Reduced Correlation Matrix: Total = 4.82403217 Average = 0.68914745				
	Eigenvalue	Difference	Proportion	Cumulative
1	4.16931551	3.39801859	0.8643	0.8643
2	0.77129692	0.58305985	0.1599	1.0242
3	0.18823707	0.17609374	0.0390	1.0632
4	0.01214333	0.06586833	0.0025	1.0657
5	-.05372500	0.03630030	-0.0111	1.0546
6	-.09002530	0.08318506	-0.0187	1.0359
7	-.17321036		-0.0359	1.0000

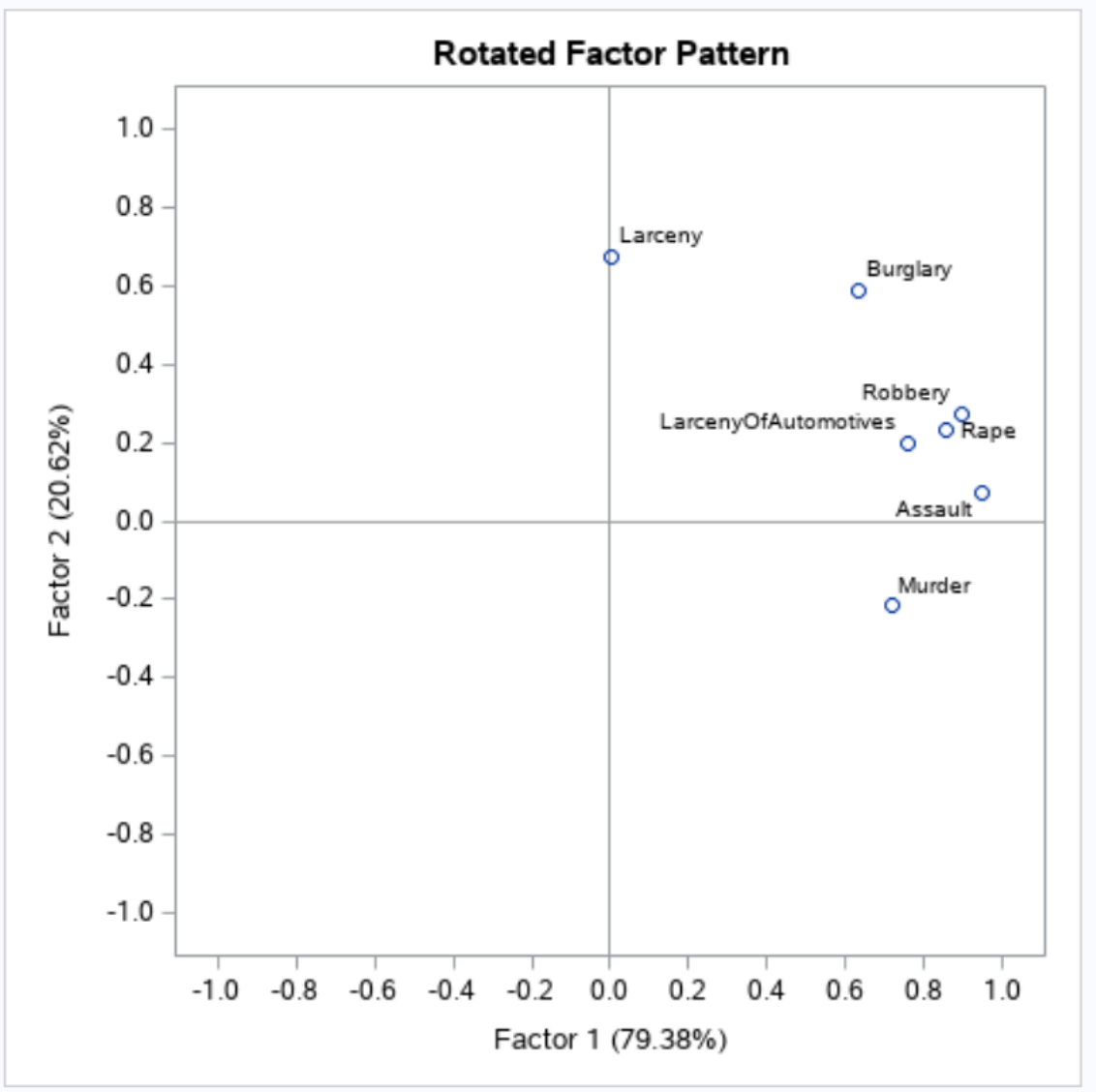
2 factors will be retained by the NFACTOR criterion.



From the scree plot above we see the natural break (also called as the elbow point) is seen at m=3. This also confirms that 2 factors are good for us to go with in this case as after this the share of variance will significantly vary for each feature.

Rotation serves to make the output more understandable, by seeking so-called "Simple Structure": A pattern of loadings where each item loads strongly on only one of the factors, and much weaker on the other factors. Rotations can be orthogonal or oblique (allowing the factors to correlate).

Varimax rotation is an orthogonal rotation of the factor axes to maximize the variance of the squared loadings of a factor (column) on all the variables (rows) in a factor matrix, which has the effect of differentiating the original variables by extracted factor. Each factor will tend to have either large or small loadings of any particular variable. A varimax solution yields results which make it as easy as possible to identify each variable with a single factor. This is the most common rotation option. However, the orthogonality (i.e., independence) of factors is often an unrealistic assumption. Oblique rotations are inclusive of orthogonal rotation, and for that reason, oblique rotations are a preferred method.



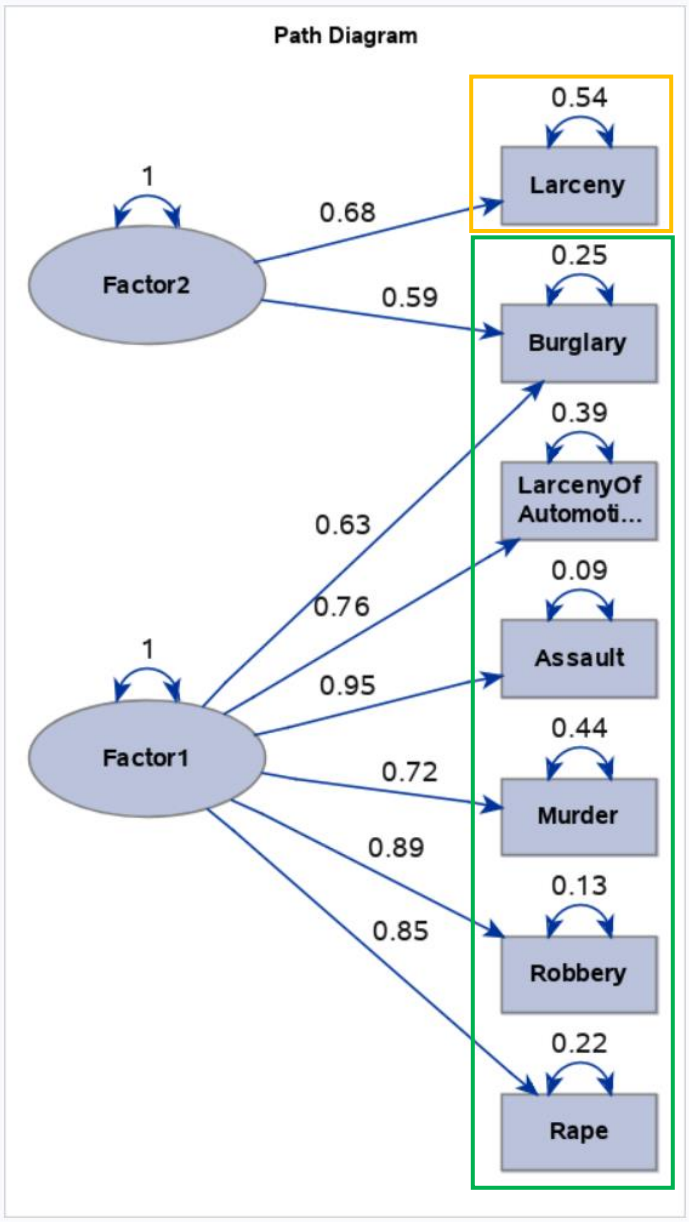
This rotated pattern simply tells us how the variance is shared across two factors. The Factor 1 on the x-axis represents the share of variance of ~80% amongst all features whereas for Factor 1 it is ~20%. As seen above, it is coming more clearly in this rotated factor pattern that 'murder' seems to be the strongest association in Factor 1 whereas for Factor 2 it is 'Larceny'. 'Burglary' looks like an exception but it is better and stronger fit to Factor 1 than Factor 2 based on out learning from the loading table above.

Path analysis is used to describe the directed dependencies among a set of variables. This includes models equivalent to any form of multiple regression analysis, factor analysis, canonical correlation analysis, discriminant analysis, as well as more general families of models in the multivariate analysis of variance and covariance analyses (MANOVA, ANOVA)

Path diagrams are like flowcharts. They show variables interconnected with lines that are used to indicate causal flow. Each path involves two variables (in either boxes or ovals) connected by either arrows (lines, usually straight, with an arrowhead on one end) or wires (lines, usually curved, with no arrowhead), or "slings" (with two arrowheads). Arrows are used to indicate "directed" relationships, or linear relationships

between two variables. An arrow from X to Y indicates a linear relationship where Y is the dependent variable and X the independent variable.

In the following path diagram one can observe how it relates in the present case of NYC felony data. There are two factors and they detail on the respective dependencies. However, we see a conflict with variable 'Burglary' which has dependency on both the factors. The factor is 0.59 for Factor 1 and 0.63 for Factor 2. They are close enough but in cases of conflict we use the range (0.3 to 0.6) and by that logic Factor 1 is a better representation of the variable 'Burglary'.



Code:

```
DATA WORK.FELONY;
        INFILE "/folders/myfolders/data/nyc_felony.dat";
        INPUT Borough $ Precinct Murder Rape Robbery Assault Burglary Larceny LarcenyOfAutomotives;
        TITLE "Project STA 9705: NYC Felony";

/* PRINCIPAL FACTOR METHOD */
PROC FACTOR METHOD=PRIN PRIORS=MAX ALL;
        VAR Murder Rape Robbery Assault Burglary Larceny LarcenyOfAutomotives;

/* PRINCIPAL FACTOR METHOD */
PROC FACTOR METHOD=PRIN PRIORS=SMC NFACT=2 ROTATE=VARIMAX PLOTS=ALL;
        VAR Murder Rape Robbery Assault Burglary Larceny LarcenyOfAutomotives;
```