Multivariate Analysis

Factor Analysis

Southern Illinois University Edwardsville

Zahra, Sagar, Steven, Yusupha

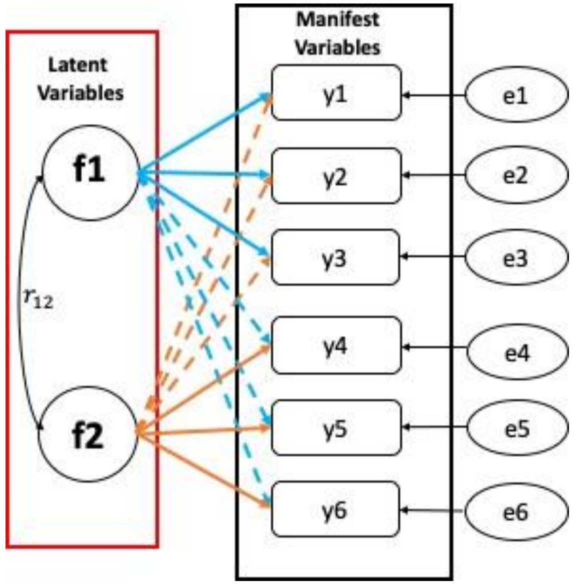
Factor Analysis & It’s Types

Factor Analysis

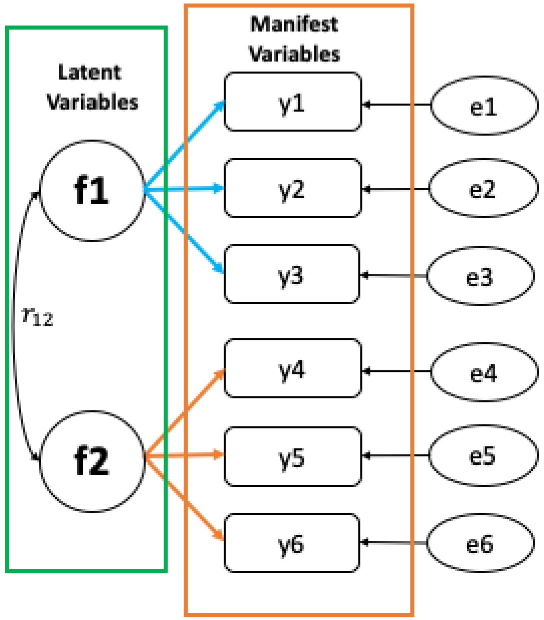
A method for *modeling observed variables, and their covariance structure,* in terms of a smaller number of *underlying unobservable (latent) “factors.”*

Exploratory

Confirmatory



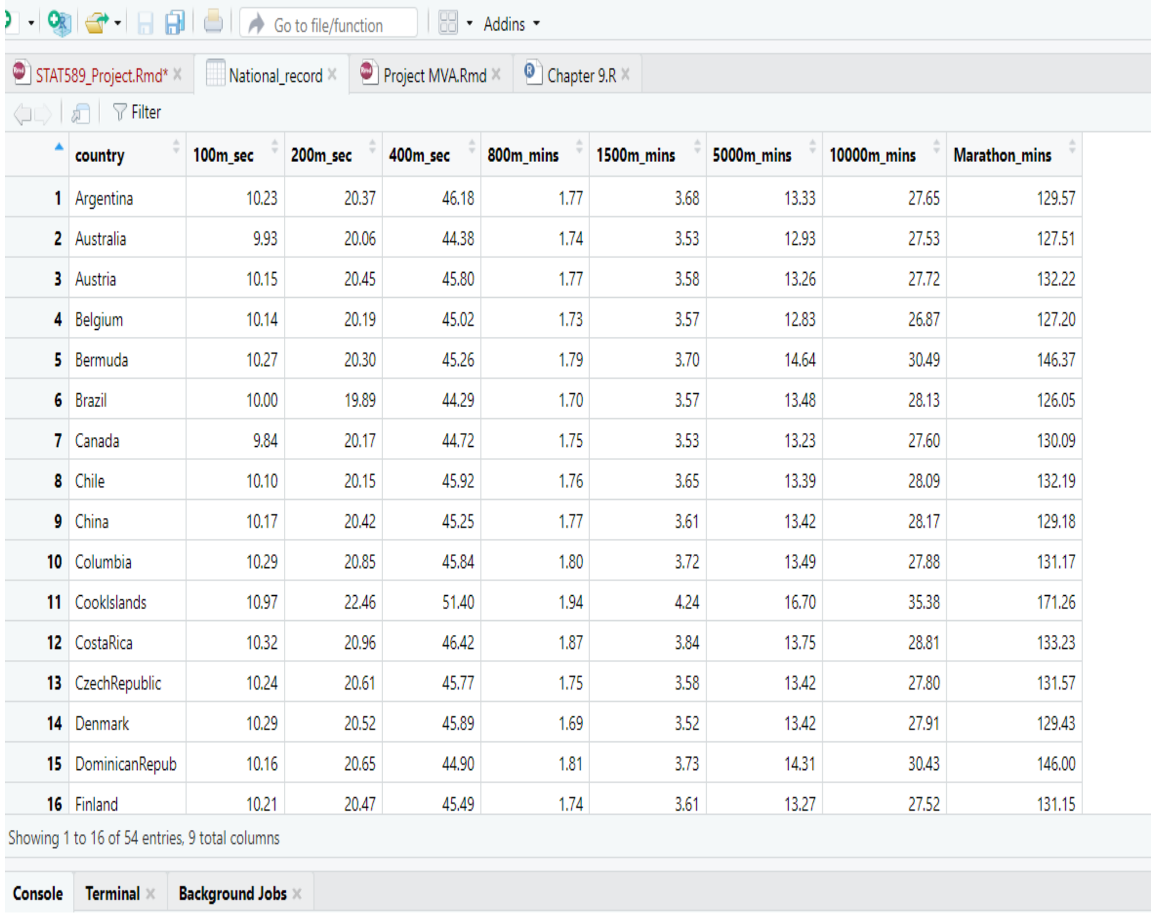
- EFA is used to *discover the factor structure of a construct* and examine its reliability.
- It is data driven.



- CFA is used to *confirm the fit of the hypothesized factor structure to observed (sample) data.*
- It is theory driven.

EFA Using R: Data and its Structure

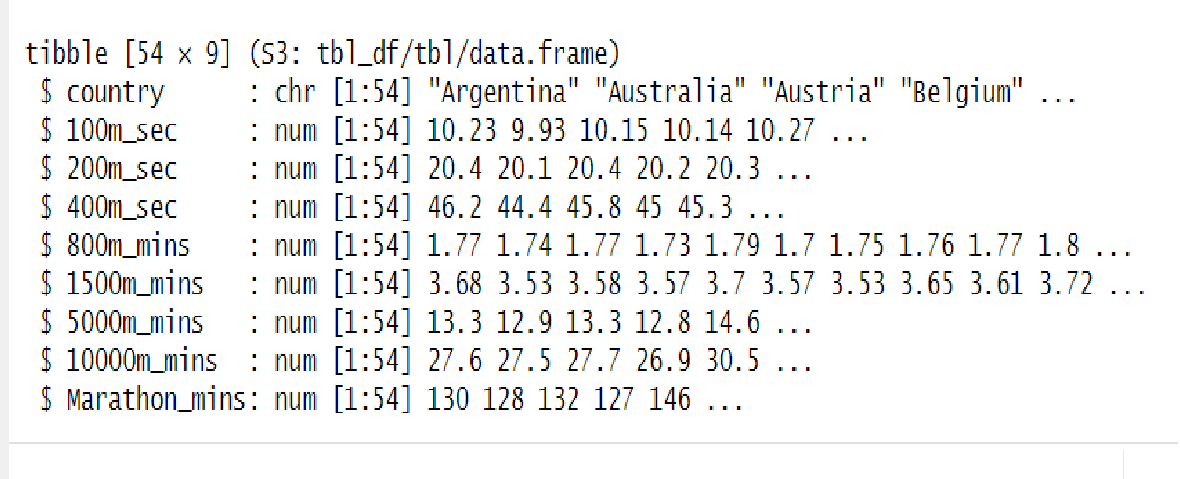
Data file → National track record.xlsx



	country	100m_sec	200m_sec	400m_sec	800m_mins	1500m_mins	5000m_mins	10000m_mins	Marathon_mins
1	Argentina	10.23	20.37	46.18	1.77	3.68	13.33	27.65	129.57
2	Australia	9.93	20.06	44.38	1.74	3.53	12.93	27.53	127.51
3	Austria	10.15	20.45	45.80	1.77	3.58	13.26	27.72	132.22
4	Belgium	10.14	20.19	45.02	1.73	3.57	12.83	26.87	127.20
5	Bermuda	10.27	20.30	45.26	1.79	3.70	14.64	30.49	146.37
6	Brazil	10.00	19.89	44.29	1.70	3.57	13.48	28.13	126.05
7	Canada	9.84	20.17	44.72	1.75	3.53	13.23	27.60	130.09
8	Chile	10.10	20.15	45.92	1.76	3.65	13.39	28.09	132.19
9	China	10.17	20.42	45.25	1.77	3.61	13.42	28.17	129.18
10	Columbia	10.29	20.85	45.84	1.80	3.72	13.49	27.88	131.17
11	CookIslands	10.97	22.46	51.40	1.94	4.24	16.70	35.38	171.26
12	CostaRica	10.32	20.96	46.42	1.87	3.84	13.75	28.81	133.23
13	CzechRepublic	10.24	20.61	45.77	1.75	3.58	13.42	27.80	131.57
14	Denmark	10.29	20.52	45.89	1.69	3.52	13.42	27.91	129.43
15	DominicanRepub	10.16	20.65	44.90	1.81	3.73	14.31	30.43	146.00
16	Finland	10.21	20.47	45.49	1.74	3.61	13.27	27.52	131.15

Showing 1 to 16 of 54 entries, 9 total columns

Data type/structure



```
tibble [54 × 9] (S3: tbl_df/tbl/data.frame)
 $ country      : chr [1:54] "Argentina" "Australia" "Austria" "Belgium" ...
 $ 100m_sec     : num [1:54] 10.23 9.93 10.15 10.14 10.27 ...
 $ 200m_sec     : num [1:54] 20.4 20.1 20.4 20.2 20.3 ...
 $ 400m_sec     : num [1:54] 46.2 44.4 45.8 45 45.3 ...
 $ 800m_mins    : num [1:54] 1.77 1.74 1.77 1.73 1.79 1.7 1.75 1.76 1.77 1.8 ...
 $ 1500m_mins   : num [1:54] 3.68 3.53 3.58 3.57 3.7 3.57 3.53 3.65 3.61 3.72 ...
 $ 5000m_mins   : num [1:54] 13.3 12.9 13.3 12.8 14.6 ...
 $ 10000m_mins  : num [1:54] 27.6 27.5 27.7 26.9 30.5 ...
 $ Marathon_mins: num [1:54] 130 128 132 127 146 ...
```

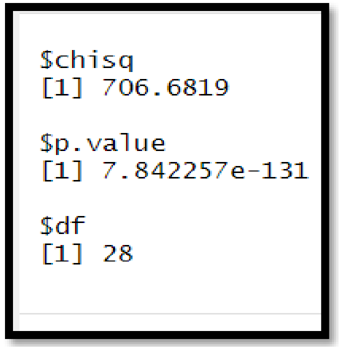
All Data columns to be analyzed are **Numeric**

EFA Using R: Initial preparation and analysis

1. **Correlation matrix:**
   - ○ Check the correlations between variables
     - ○ There are essentially two potential problems:
       1. Correlations that are not high enough
       2. Correlations that are too high.

Bartlett's Test of Sphericity:

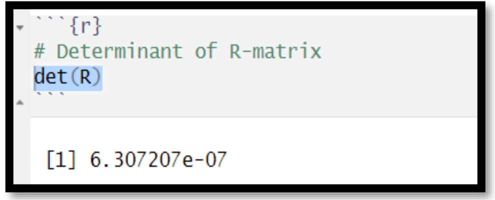
- ○ Compares an observed correlation matrix to the identity matrix
  - ○ H_0: R Matrix = Identity Matrix (i.e., There is No Correlation Between Variables)
  - ○ H_1: R Matrix ≠ Identity Matrix (i.e., There is a Correlation Between Variables)



```
$chisq
[1] 706.6819

$p.value
[1] 7.842257e-131

$df
[1] 28
```

- ▪ P-Value < 0.01; so, Bartlett's test is highly significant (i.e., *R*-matrix is not an identity matrix); $\chi_{2}(253) = 706.6819$, $p < .01$, and therefore factor analysis is appropriate.

The Determinant of the *R*-matrix should be greater than 0.00001



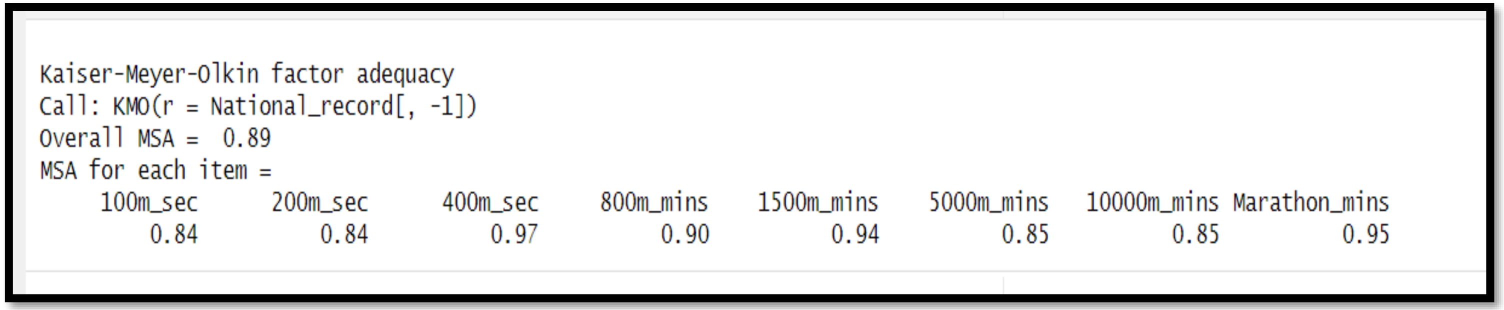
~~~
```{r}
# Determinant of R-matrix
det(R)
```

[1] 6.307207e-07
~~~

EFA Using R: Initial preparation and analysis…

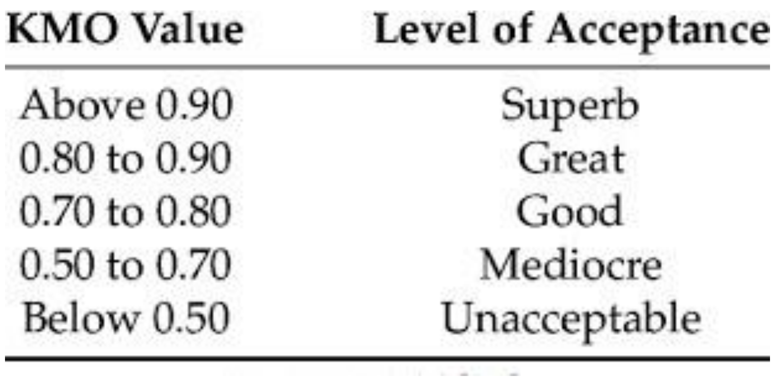
3. **Sample Size:**
   - KMO Test (Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy):

KMO → The ratio of the squared correlation between variables to the squared partial correlation between variables.

- The KMO statistic varies between 0 and 1.
- KMO close to 0 indicates diffusion in the pattern of correlations → factor analysis is likely to be inappropriate
- KMO close to 1 indicates that patterns of correlations are relatively compact → factor analysis should yield distinct and reliable factors.



```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = National_record[, -1])
Overall MSA =  0.89
MSA for each item = 
     100m_sec      200m_sec      400m_sec      800m_mins     1500m_mins    5000m_mins   10000m_mins Marathon_mins 
         0.84          0.84          0.97          0.90          0.94          0.85          0.85          0.95 
```



KMO Value	Level of Acceptance
Above 0.90	Superb
0.80 to 0.90	Great
0.70 to 0.80	Good
0.50 to 0.70	Mediocre
Below 0.50	Unacceptable

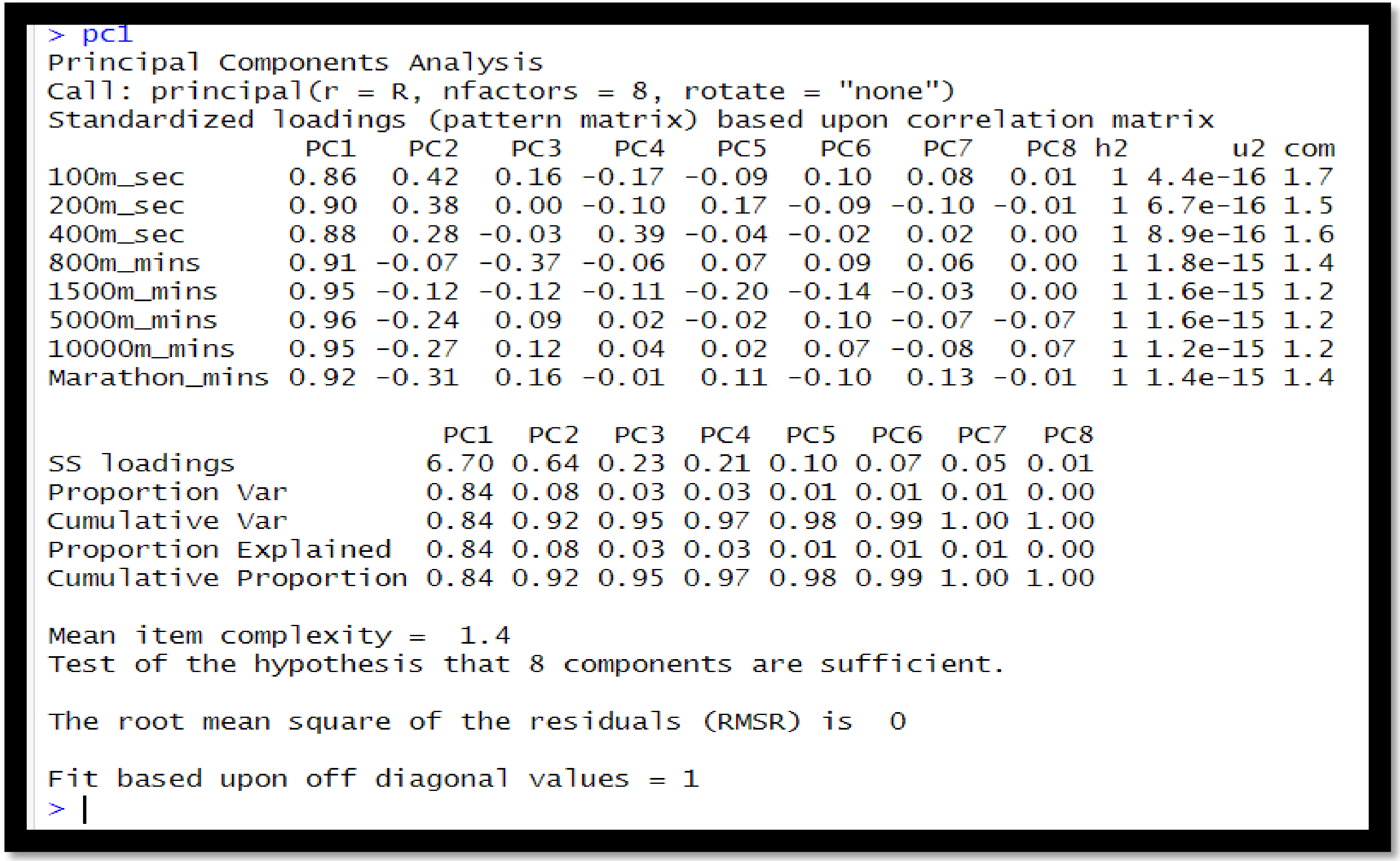
- Because Overall KMO = 0.89 & values of KMO for all the variables greater than 0.5 so, sample size and the data are adequate for factor analysis
- If any variables with KMO values below .5 then you should consider excluding them from the analysis.

EFA Using R: Factor extraction

- **Methods of Factor Extraction**
  - PCA Method (Scree plot)
  - Maximum Likelihood

EFA Using R: Factor extraction

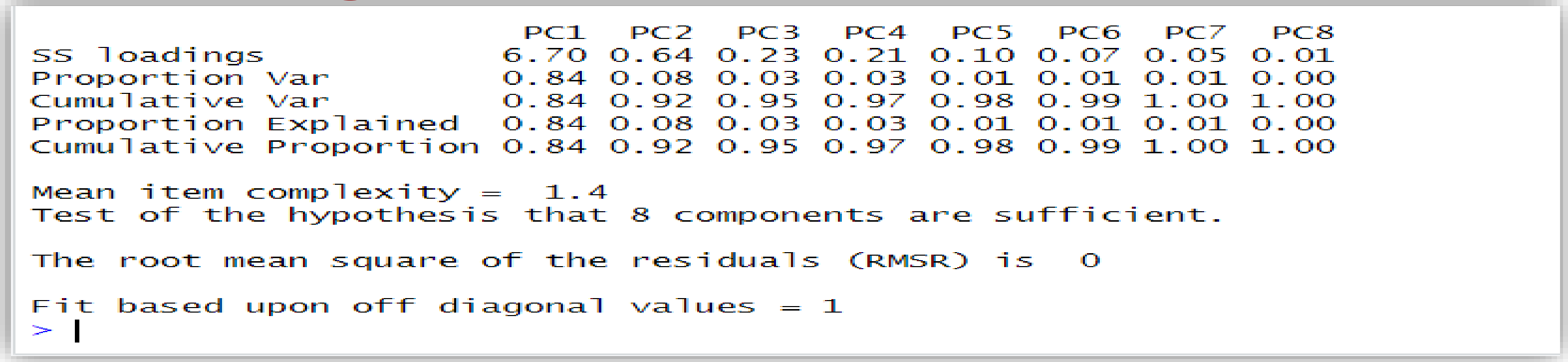
Factor Loading Matrix



```
> pc1
Principal Components Analysis
Call: principal(r = R, nfactors = 8, rotate = "none")
Standardized loadings (pattern matrix) based upon correlation matrix
              PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8 h2      u2 com
100m_sec     0.86  0.42  0.16 -0.17 -0.09  0.10  0.08  0.01  1 4.4e-16 1.7
200m_sec     0.90  0.38  0.00 -0.10  0.17 -0.09 -0.10 -0.01  1 6.7e-16 1.5
400m_sec     0.88  0.28 -0.03  0.39 -0.04 -0.02  0.02  0.00  1 8.9e-16 1.6
800m_mins    0.91 -0.07 -0.37 -0.06  0.07  0.09  0.06  0.00  1 1.8e-15 1.4
1500m_mins   0.95 -0.12 -0.12 -0.11 -0.20 -0.14 -0.03  0.00  1 1.6e-15 1.2
5000m_mins   0.96 -0.24  0.09  0.02 -0.02  0.10 -0.07 -0.07  1 1.6e-15 1.2
10000m_mins  0.95 -0.27  0.12  0.04  0.02  0.07 -0.08  0.07  1 1.2e-15 1.2
Marathon_mins 0.92 -0.31  0.16 -0.01  0.11 -0.10  0.13 -0.01  1 1.4e-15 1.4

                       PC1  PC2  PC3  PC4  PC5  PC6  PC7  PC8
SS loadings           6.70 0.64 0.23 0.21 0.10 0.07 0.05 0.01
Proportion Var        0.84 0.08 0.03 0.03 0.01 0.01 0.01 0.00
Cumulative Var        0.84 0.92 0.95 0.97 0.98 0.99 1.00 1.00
Proportion Explained  0.84 0.08 0.03 0.03 0.01 0.01 0.01 0.00
Cumulative Proportion 0.84 0.92 0.95 0.97 0.98 0.99 1.00 1.00

Mean item complexity =  1.4
Test of the hypothesis that 8 components are sufficient.

The root mean square of the residuals (RMSR) is  0

Fit based upon off diagonal values = 1
> |
```

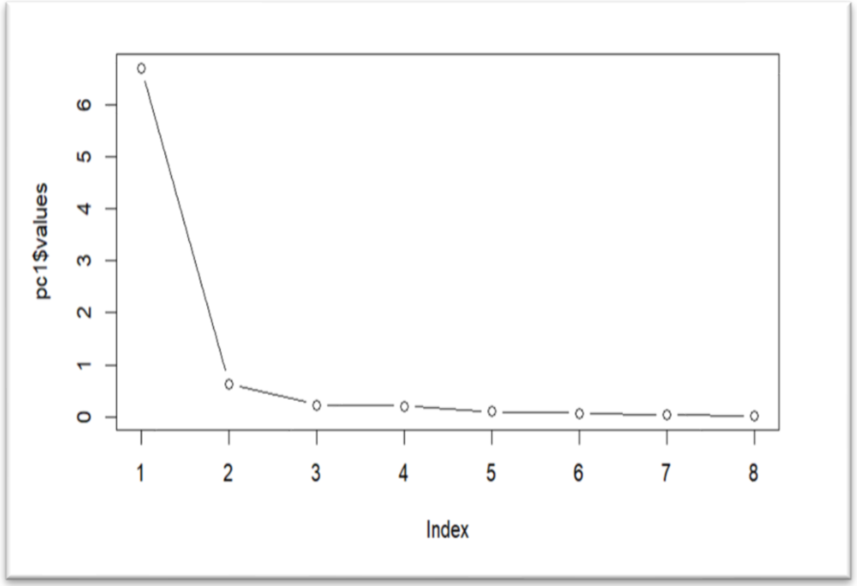
- **PCA Method:**
  - By extracting the factors, inspect two columns, communality(h2) and uniqueness(u2)
    - h2 (communalities) → *All equal to 1 (Explained all of the variance in every variable)*
      - Because Factor extracted 8 = number of variables
      - When we extract fewer factors (or components) we’ll have lower communalities.
    - u2 (amount of unique variance for each variable) → (1 – communalities) → all of the uniqueness's are 0

EFA Using R: Factor Extraction



```
                        PC1  PC2  PC3  PC4  PC5  PC6  PC7  PC8
SS loadings            6.70 0.64 0.23 0.21 0.10 0.07 0.05 0.01
Proportion Var         0.84 0.08 0.03 0.01 0.01 0.01 0.01 0.00
Cumulative Var         0.84 0.92 0.95 0.97 0.98 0.99 1.00 1.00
Proportion Explained   0.84 0.08 0.03 0.03 0.01 0.01 0.01 0.00
Cumulative Proportion  0.84 0.92 0.95 0.97 0.98 0.99 1.00 1.00

Mean item complexity =  1.4
Test of the hypothesis that 8 components are sufficient.

The root mean square of the residuals (RMSR) is  0

Fit based upon off diagonal values = 1
> 
```

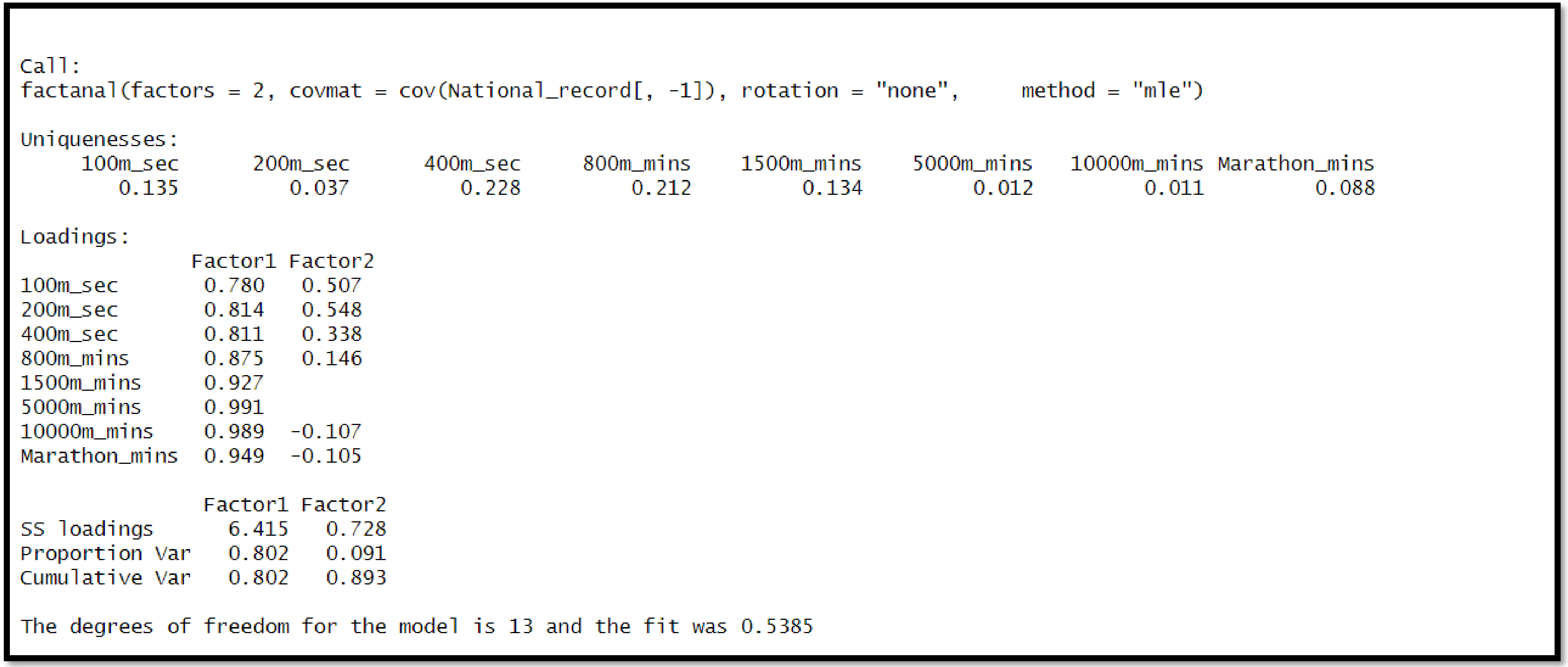
- **Eigen Values:**
  - The eigenvalues associated with each factor represent the *variance explained by that particular linear component*.
  - **R** calls these SS loadings (sums of squared loadings)
    - **Factor 1 →** explains 6.70 units of variance out of a possible 8 (the number of factors) so as a proportion this is 6.70/8 = 0.8375 ≈ 8.4; so, factor 1 explains 84% of the total variance.



- **Factor Extraction Criteria:**
  - According to Kaiser's criterion (eigenvalues > 1) → We can pick one components (or factors)
  - We should also consider the ***scree plot***.
  - By Scree plot we should take 2 components (or Factors)

The evidence from the scree plot and formula test(later slide) suggests a two-component solution may be the best.

EFA Using R: Factor Extraction



```
Call:
factanal(factors = 2, covmat = cov(National_record[, -1]), rotation = "none",     method = "mle")

Uniquenesses:
     100m_sec      200m_sec      400m_sec     800m_mins    1500m_mins    5000m_mins   10000m_mins Marathon_mins
        0.135         0.037         0.228         0.212         0.134         0.012         0.011         0.088

Loadings:
              Factor1 Factor2
100m_sec       0.780   0.507
200m_sec       0.814   0.548
400m_sec       0.811   0.338
800m_mins      0.875   0.146
1500m_mins     0.927
5000m_mins     0.991
10000m_mins    0.989  -0.107
Marathon_mins  0.949  -0.105

               Factor1 Factor2
SS loadings      6.415   0.728
Proportion Var   0.802   0.091
Cumulative Var   0.802   0.893

The degrees of freedom for the model is 13 and the fit was 0.5385
```

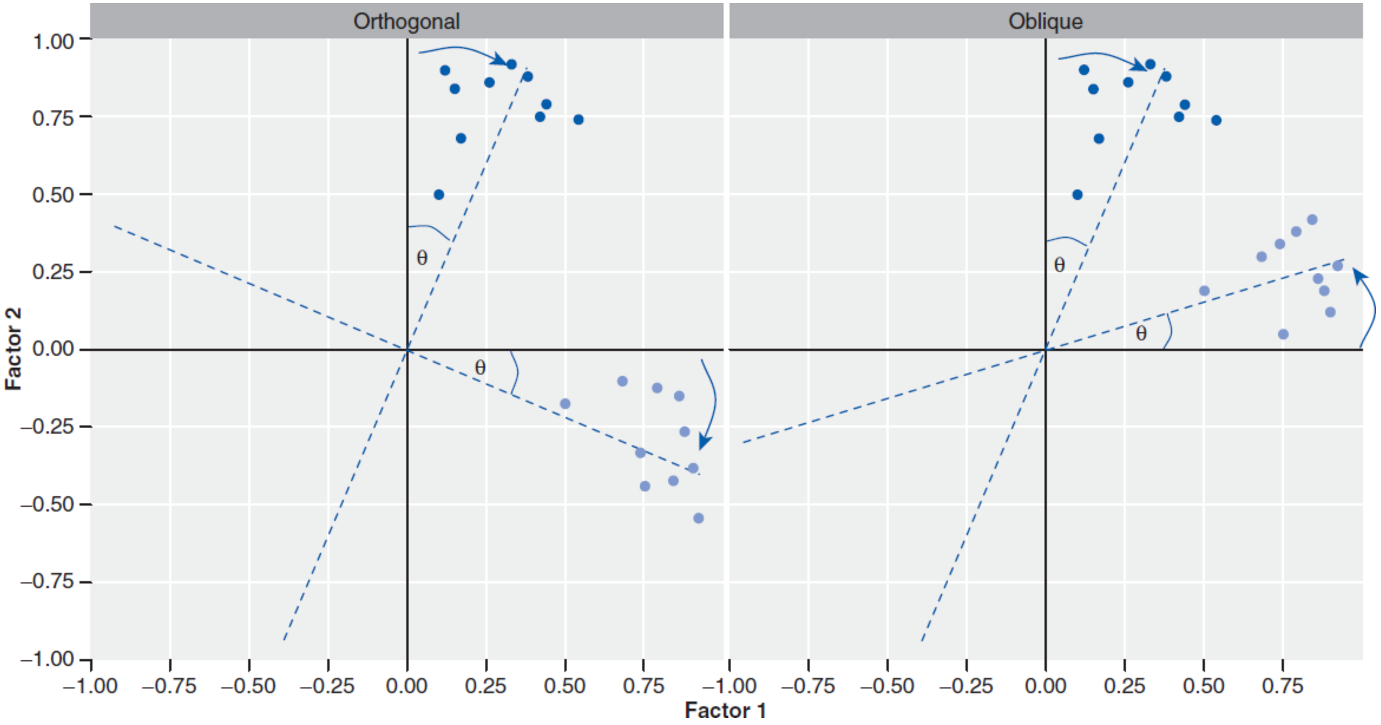
- To decide **how many factors to extract**:
  - We look at maximum likelihood approach and the scree plot.

Formal test for the number of factors -maximum likelihood approach

EFA Using R: Factor Rotation

After the factors extraction:

- **Factor Loading:** Calculate to what degree variables load on these factors (i.e., calculate the loading of the variable on each factor).
  - Generally, most variables have *high loadings on the most important factor* and *small loadings on all other factors*.



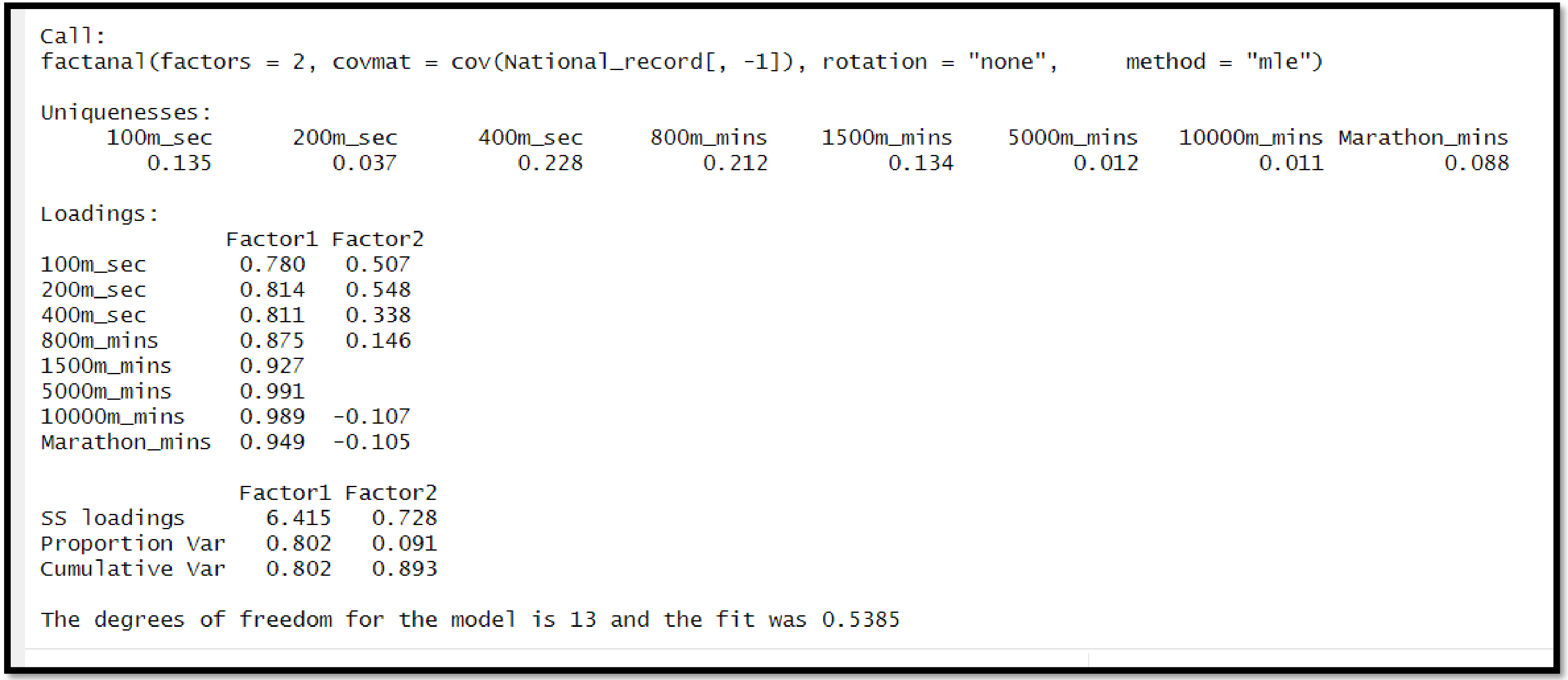
- *Orthogonal rotation ensures that the factors remain independent or uncorrelated (perpendicular).*
- *Oblique rotation allow the factors to correlate (hence, do not perpendicular).*

Rotation maximizes the loading of each variable on one of the extracted factors while minimizing the loading on all other factors.

EFA Using R: Factor Rotation → IF None

- - The factor loading for last 4 variables seems to contribute for Factor-1
- - The first factor is dominated by short length race time, so we can label this factor as **Speed**

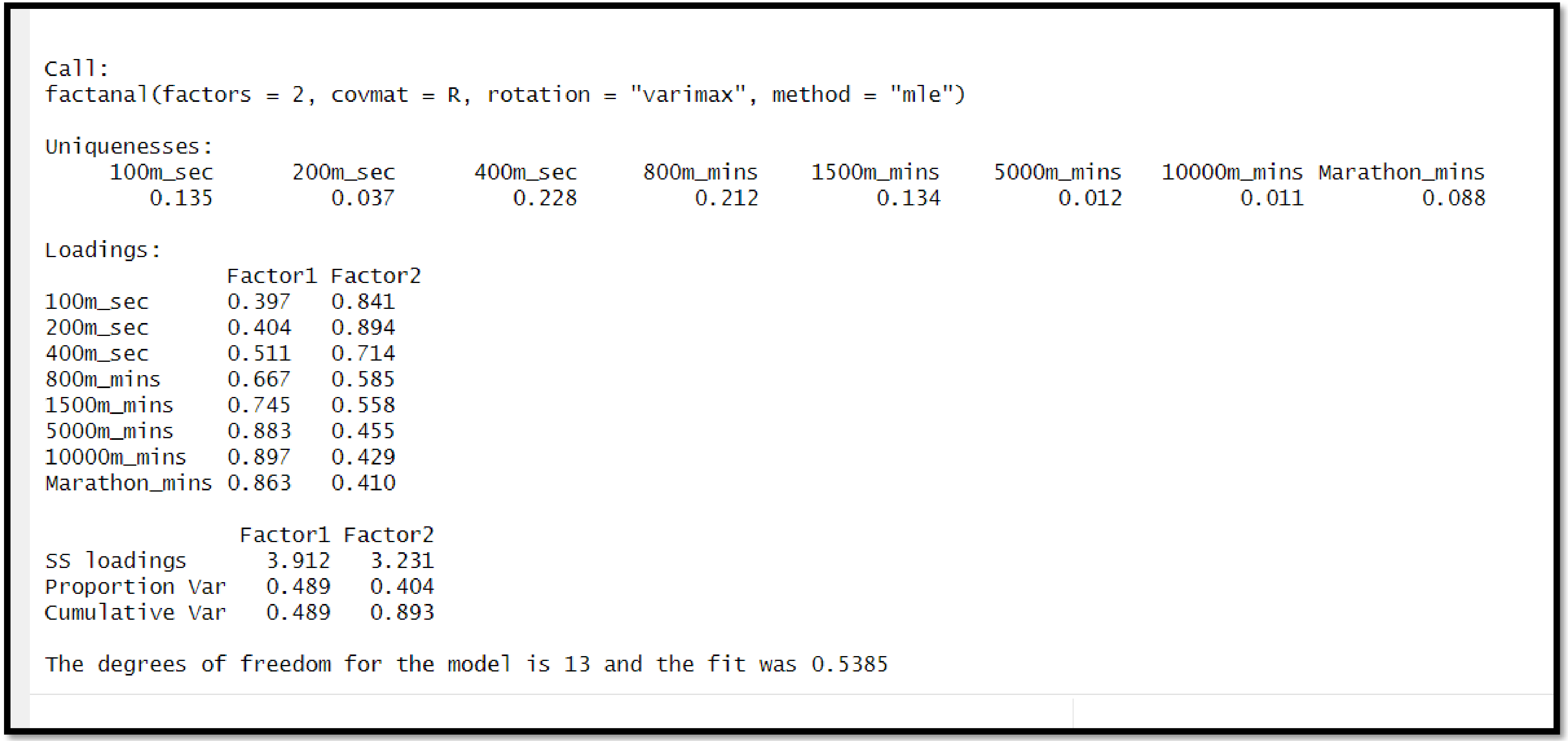
- - The factor loading for first 4 variables seems to contribute for Factor-2
- - The second factor is dominated by long length race time, so we can label this factor as **endurance**



```
Call:
factanal(factors = 2, covmat = cov(National_record[, -1]), rotation = "none",     method = "mle")

Uniquenesses:
     100m_sec      200m_sec      400m_sec     800m_mins     1500m_mins     5000m_mins    10000m_mins Marathon_mins
        0.135         0.037         0.228         0.212          0.134          0.012          0.011         0.088

Loadings:
              Factor1 Factor2
100m_sec       0.780   0.507 
200m_sec       0.814   0.548 
400m_sec       0.811   0.338 
800m_mins      0.875   0.146 
1500m_mins     0.927         
5000m_mins     0.991         
10000m_mins    0.989  -0.107 
Marathon_mins  0.949  -0.105 

               Factor1 Factor2
SS loadings      6.415   0.728
Proportion Var   0.802   0.091
Cumulative Var   0.802   0.893

The degrees of freedom for the model is 13 and the fit was 0.5385
```

- - First factor explains 80.2% of Total variance in our dataset
- -Second factor explains 9.1% of Total variance in our dataset

EFA Using R: Factor Rotation → Orthogonal rotation (varimax)

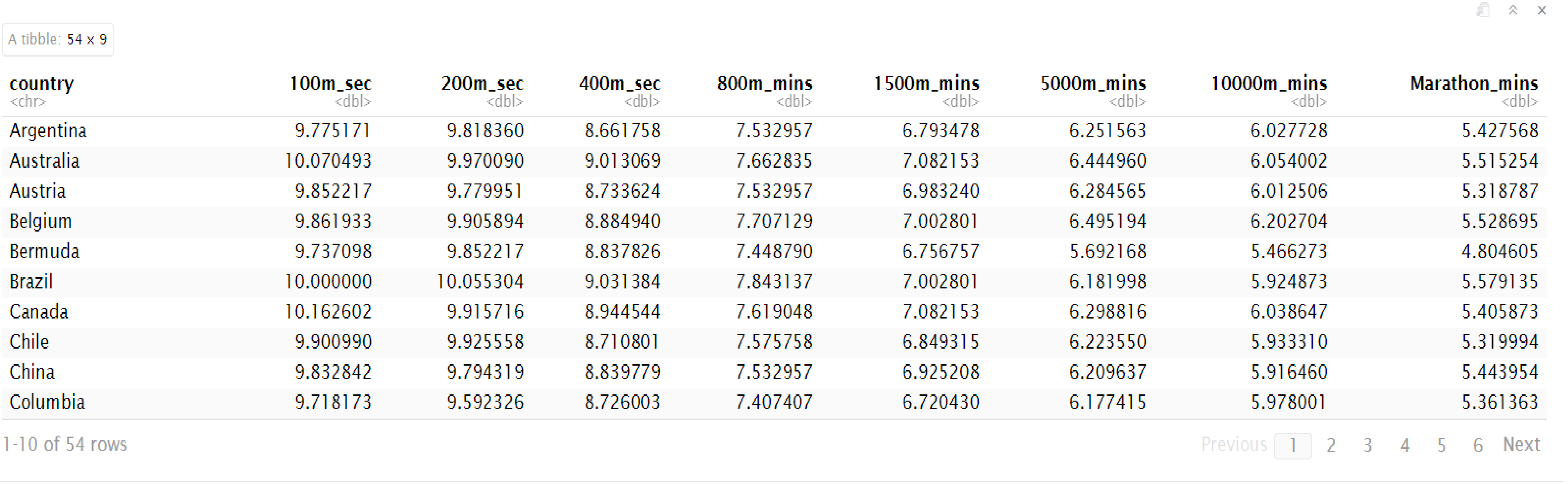


```
Call:
factanal(factors = 2, covmat = R, rotation = "varimax", method = "mle")

Uniquenesses:
     100m_sec      200m_sec      400m_sec     800m_mins    1500m_mins    5000m_mins   10000m_mins Marathon_mins 
        0.135         0.037         0.228         0.212         0.134         0.012         0.011         0.088 

Loadings:
              Factor1 Factor2
100m_sec      0.397   0.841  
200m_sec      0.404   0.894  
400m_sec      0.511   0.714  
800m_mins     0.667   0.585  
1500m_mins    0.745   0.558  
5000m_mins    0.883   0.455  
10000m_mins   0.897   0.429  
Marathon_mins 0.863   0.410  

               Factor1 Factor2
SS loadings      3.912   3.231
Proportion Var   0.489   0.404
Cumulative Var   0.489   0.893

The degrees of freedom for the model is 13 and the fit was 0.5385
```

After applying the Orthogonal rotation.

- - First factor explains 48.9% of Total variance in our dataset
- -Second factor explains 40.4% of Total variance in our dataset

EFA Using R:Data Conversion



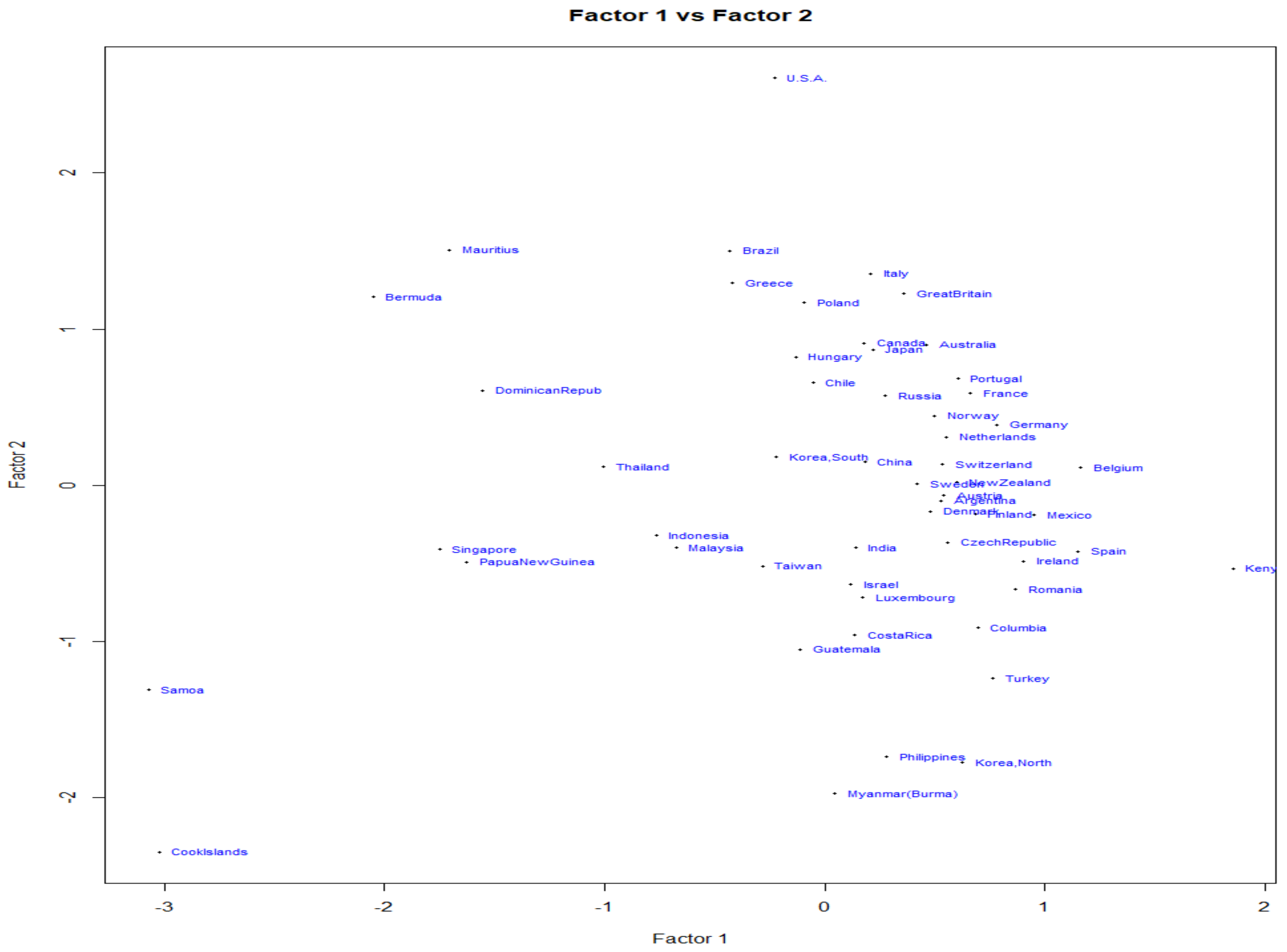
A tibble: 54 x 9

country	100m_sec	200m_sec	400m_sec	800m_mins	1500m_mins	5000m_mins	10000m_mins	Marathon_mins
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Argentina	9.775171	9.818360	8.661758	7.532957	6.793478	6.251563	6.027728	5.427568
Australia	10.070493	9.970090	9.013069	7.662835	7.082153	6.444960	6.054002	5.515254
Austria	9.852217	9.779951	8.733624	7.532957	6.983240	6.284565	6.012506	5.318787
Belgium	9.861933	9.905894	8.884940	7.707129	7.002801	6.495194	6.202704	5.528695
Bermuda	9.737098	9.852217	8.837826	7.448790	6.756757	5.692168	5.466273	4.804605
Brazil	10.000000	10.055304	9.031384	7.843137	7.002801	6.181998	5.924873	5.579135
Canada	10.162602	9.915716	8.944544	7.619048	7.082153	6.298816	6.038647	5.405873
Chile	9.900990	9.925558	8.710801	7.575758	6.849315	6.223550	5.933310	5.319994
China	9.832842	9.794319	8.839779	7.532957	6.925208	6.209637	5.916460	5.443954
Columbia	9.718173	9.592326	8.726003	7.407407	6.720430	6.177415	5.978001	5.361363

1-10 of 54 rows

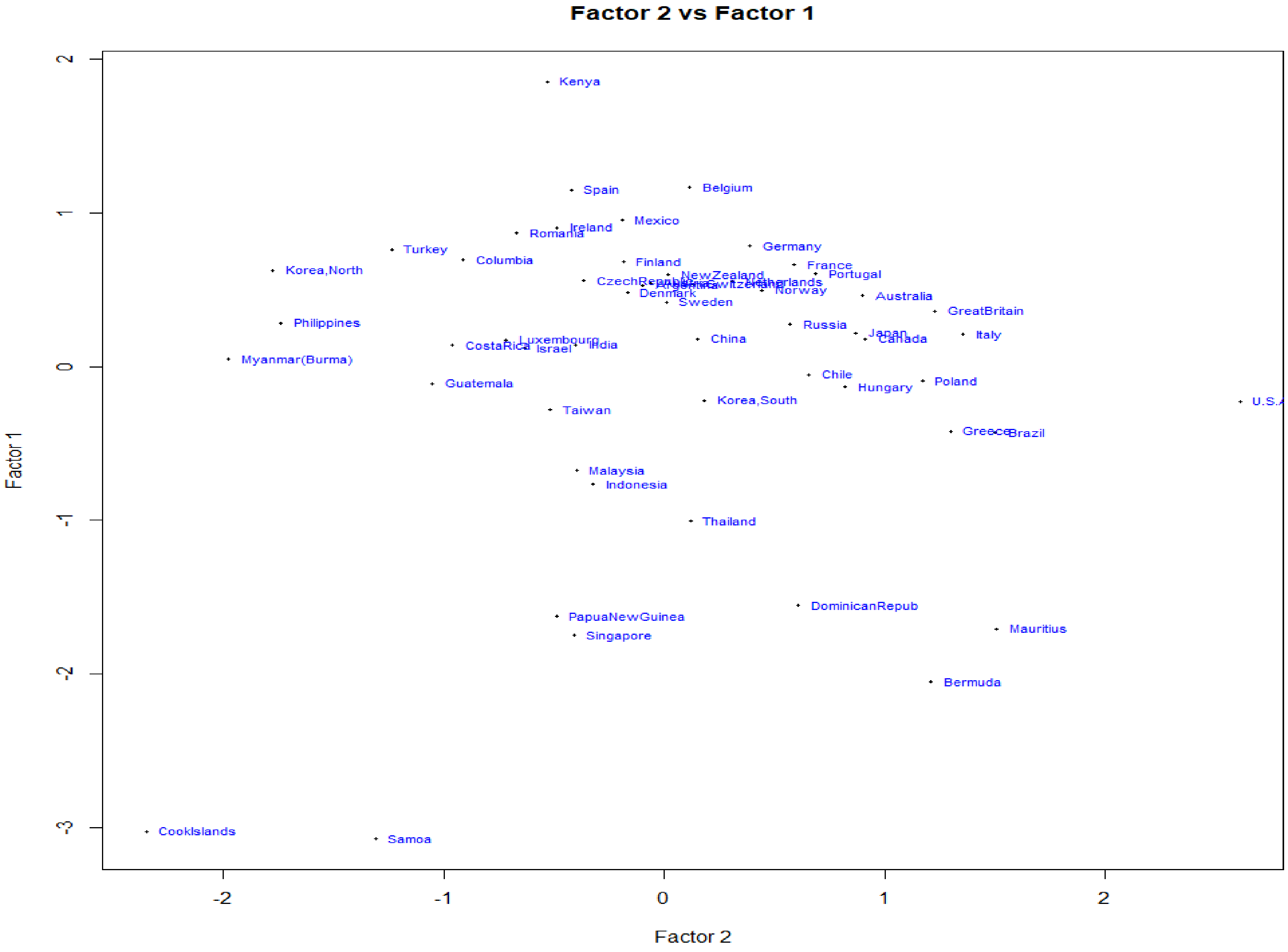
EFA Using R: Factor plot → Factor 1 vs Factor 2

Interpretation:-

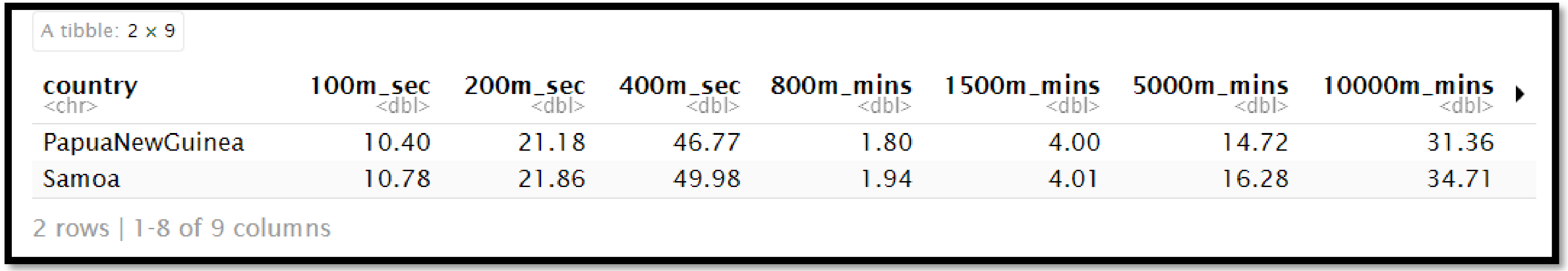


EFA Using R: Factor plot → Factor 1 vs Factor 2

Interpretation:-



EFA Using R: Outlier detection and Removal

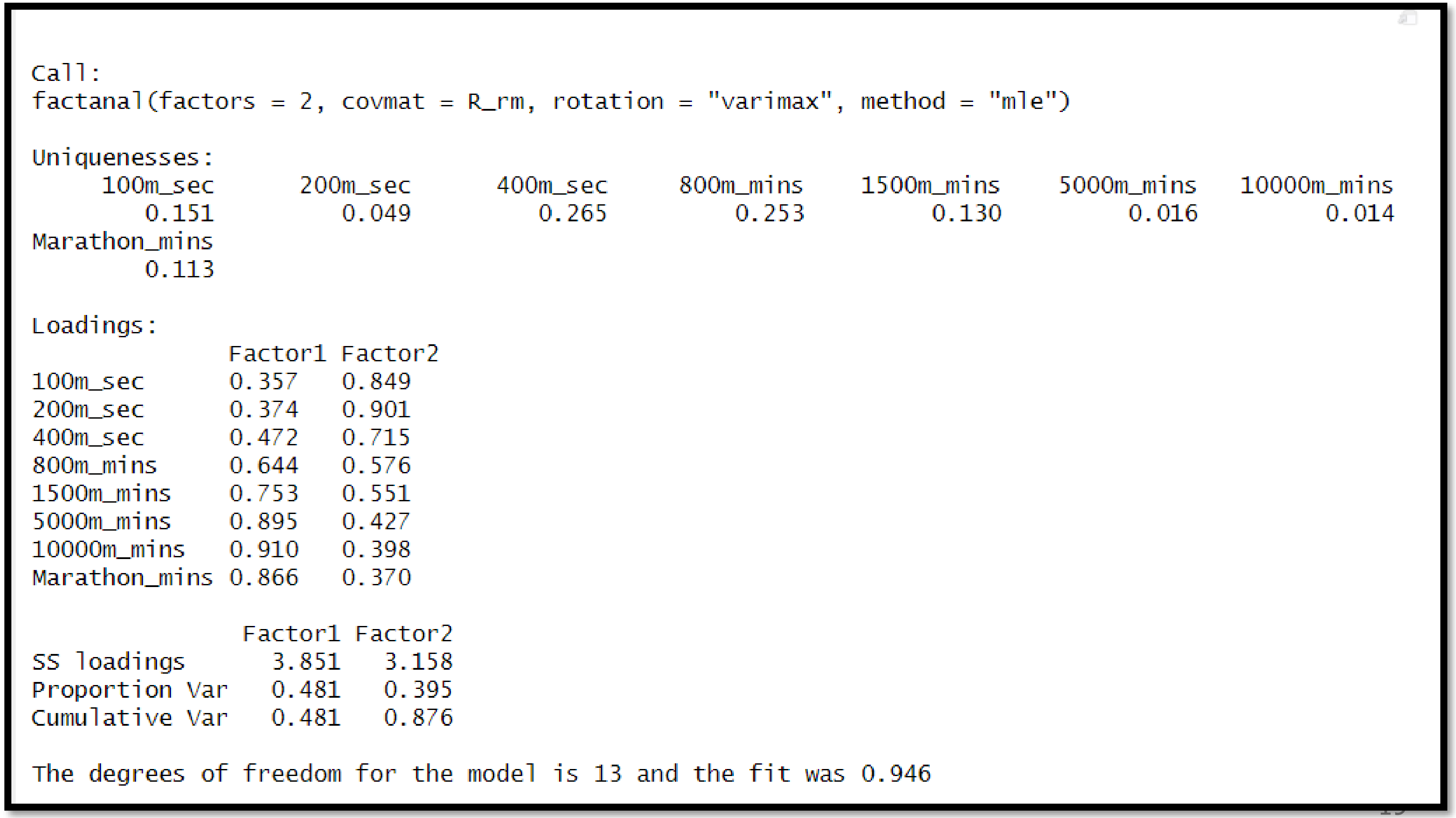


A tibble: 2 × 9

country	100m_sec	200m_sec	400m_sec	800m_mins	1500m_mins	5000m_mins	10000m_mins
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
PapuaNewGuinea	10.40	21.18	46.77	1.80	4.00	14.72	31.36
Samoa	10.78	21.86	49.98	1.94	4.01	16.28	34.71

2 rows | 1-8 of 9 columns

These are the two observation
Which act as outlier
when observer jointly

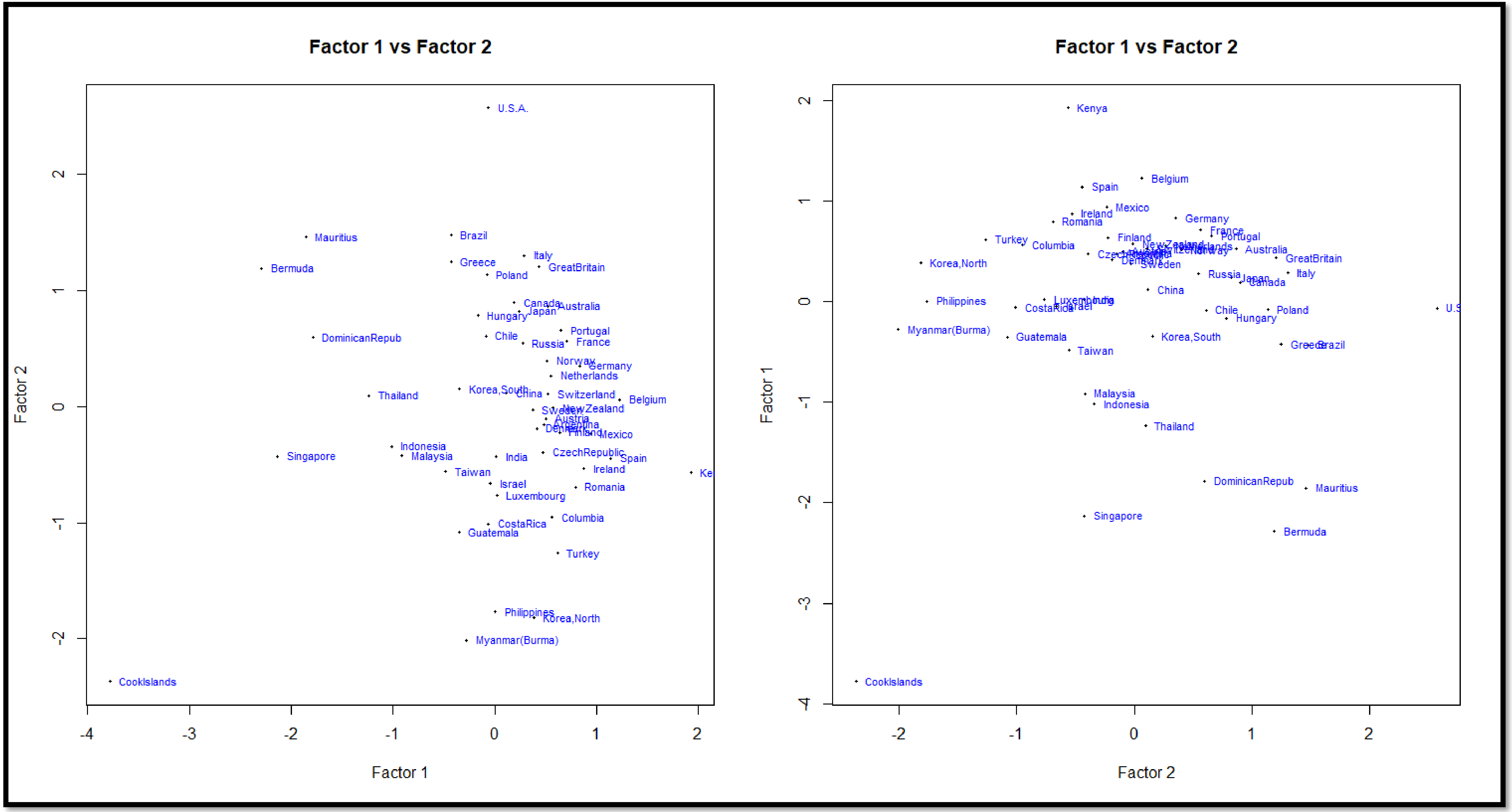


```
Call:
factanal(factors = 2, covmat = R_rm, rotation = "varimax", method = "mle")

Uniquenesses:
     100m_sec      200m_sec      400m_sec     800m_mins    1500m_mins    5000m_mins   10000m_mins
        0.151         0.049         0.265         0.253         0.130         0.016         0.014
Marathon_mins
        0.113

Loadings:
              Factor1 Factor2
100m_sec      0.357   0.849
200m_sec      0.374   0.901
400m_sec      0.472   0.715
800m_mins     0.644   0.576
1500m_mins    0.753   0.551
5000m_mins    0.895   0.427
10000m_mins   0.910   0.398
Marathon_mins 0.866   0.370

               Factor1 Factor2
SS loadings      3.851   3.158
Proportion Var   0.481   0.395
Cumulative Var   0.481   0.876

The degrees of freedom for the model is 13 and the fit was 0.946
```

We are getting the same two Factor but this time with each variable having good factor loading.

Outlier detection make interpretation and Factor selection nicer.

EFA Using R: Factor plot → Factor 1 vs Factor 2 (W/O Outlier)



QUESTIONS