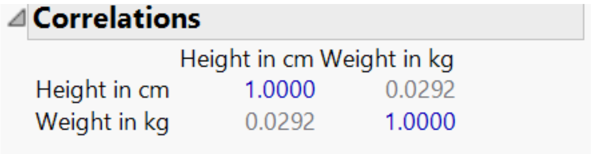
“The work contained and presented here is my work and my work alone.”

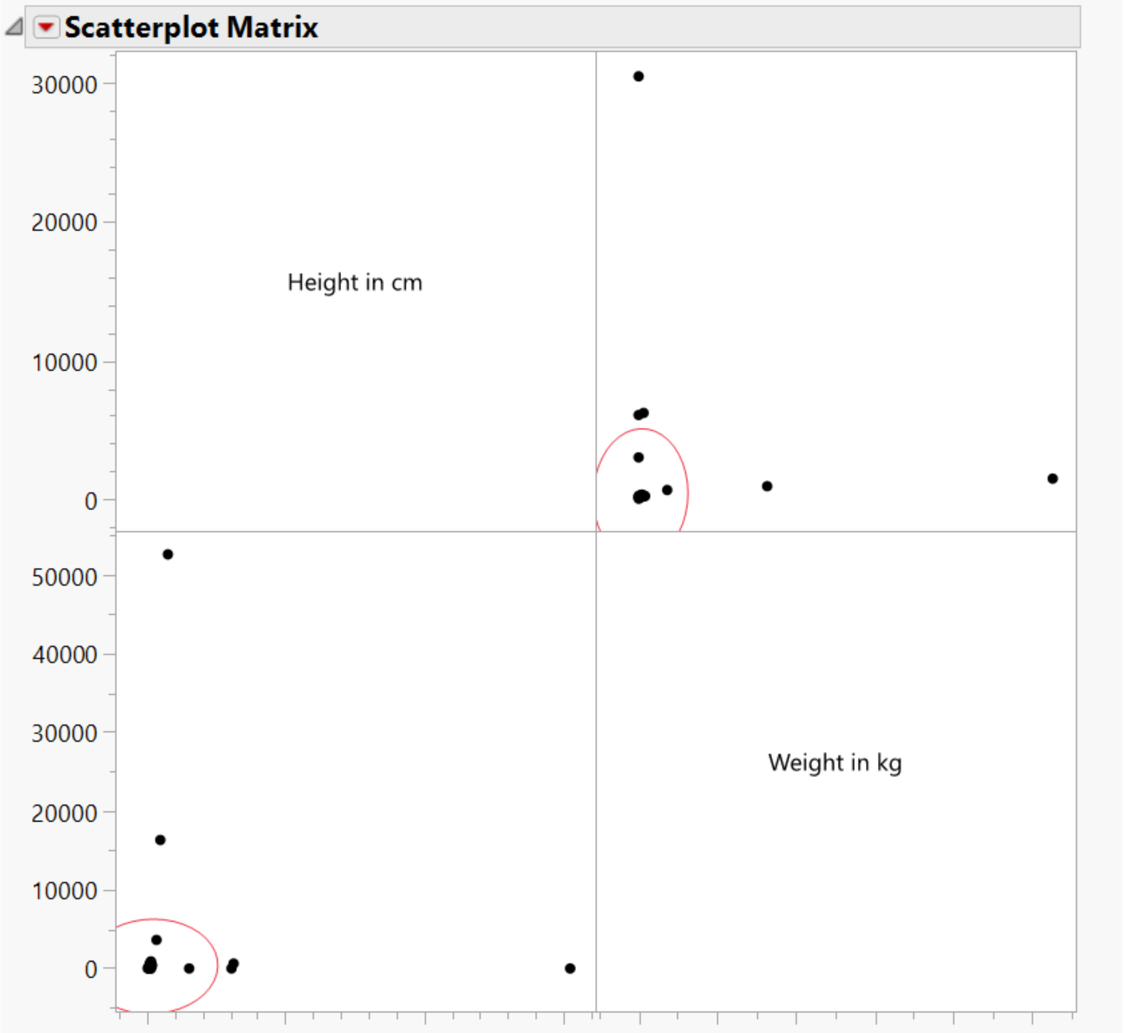
1. Create a scatterplot matrix for Height in cm and Weight in kg. What is the correlation of the two variables? Include a screen shot of the scatterplot matrix.

Sol:- The correlation is very weak positive correlation.



Correlations

	Height in cm	Weight in kg
Height in cm	1.0000	0.0292
Weight in kg	0.0292	1.0000



2. Look at the Mahalanobis Distances and exclude the eight most distant rows. (Don’t hide them so you can tell these rows apart from the 241 hidden and excluded rows that you started with.) Which characters did you exclude?

Sol :- The characters that were excluded are :-

32 -Anti Monitor

“The work contained and presented here is my work and my work alone.”

124- Bloodwraith

255 – Fin Fanf foom

282 – Giganta

308 – Groot

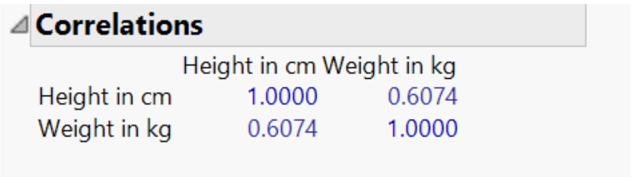
655- Surtur

693- Utgard-Loki

739- Ymir

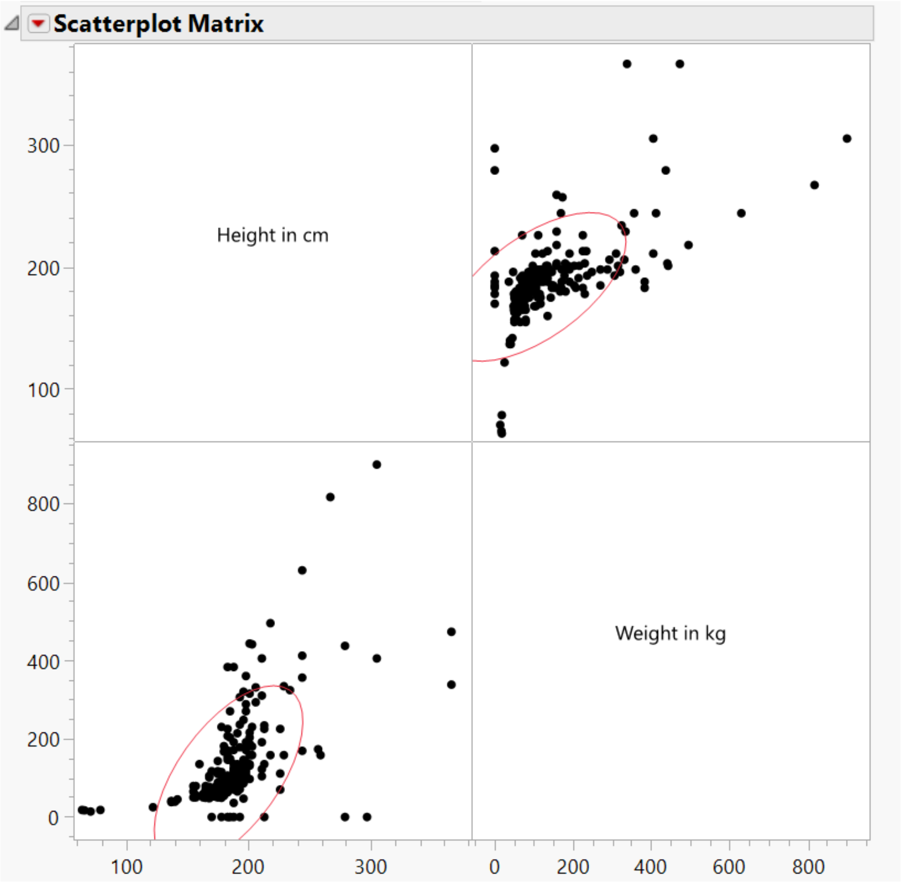
3. Recreate the scatterplot matrix for Height in cm and Weight in kg. How did the correlation change? Include a screen shot of the scatterplot matrix.

Sol:- Since the correlation is 0.6074 (i.e ranges between 0.5 -0.7) and is in moderate blue color, the correlation has changed to a Strong positive correlation. Since we have excluded the 8 most distant rows.



Correlations

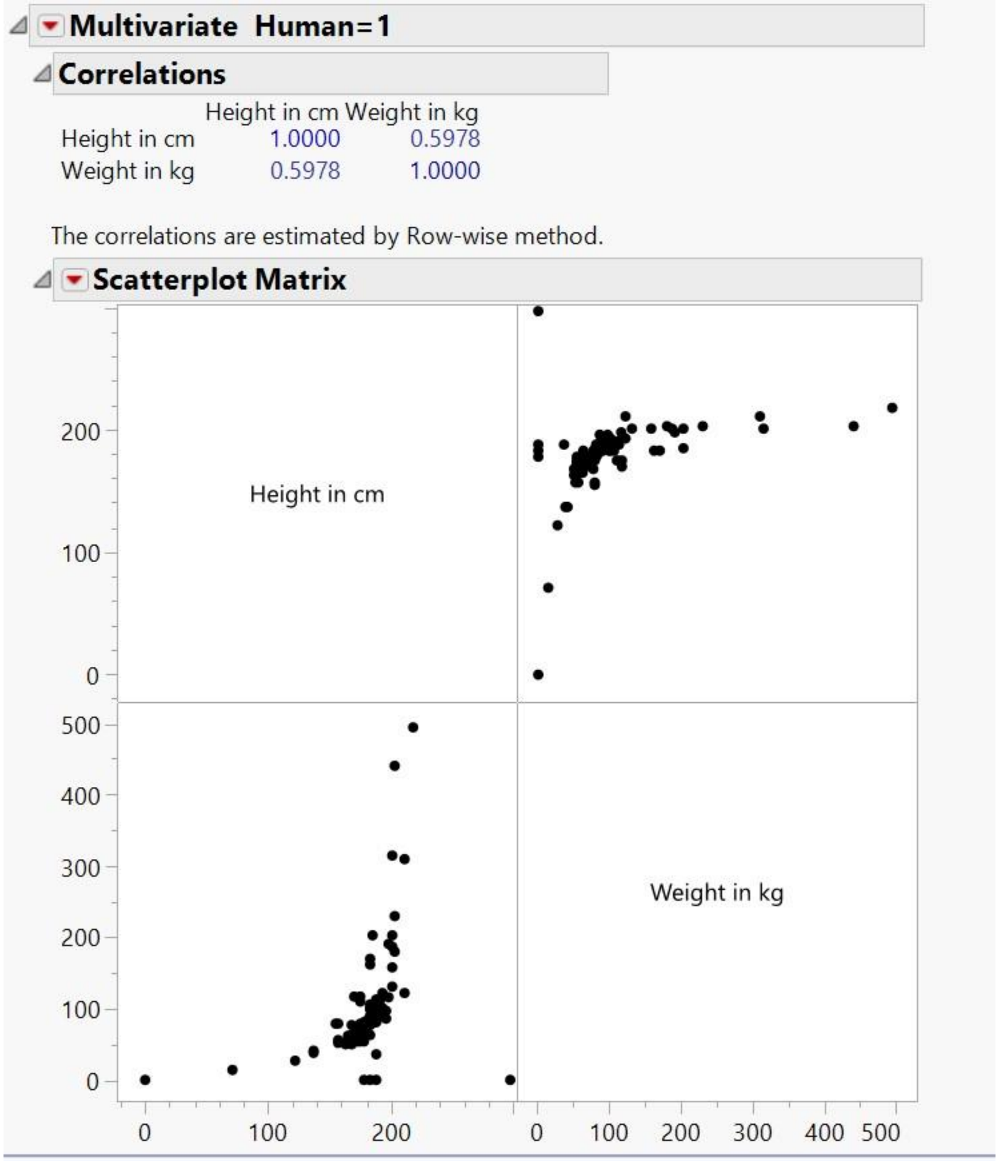
	Height in cm	Weight in kg
Height in cm	1.0000	0.6074
Weight in kg	0.6074	1.0000



“The work contained and presented here is my work and my work alone.”

4. What have these four questions taught you about how outliers impact the data?

Sol:- Outliers are just observations that were not following the same pattern than the other ones. (Like the 8 characters that we excluded in Q2) Outliers influence the data on a whole, they create negative bias & anomalies in the data. The correlation of the 2 variables was weaker when the outliers were included, and stronger when we excluded them, in fact the outliers are the disturbance to the analysis as well at times it is what we are looking for.



Multivariate Human=1

Correlations

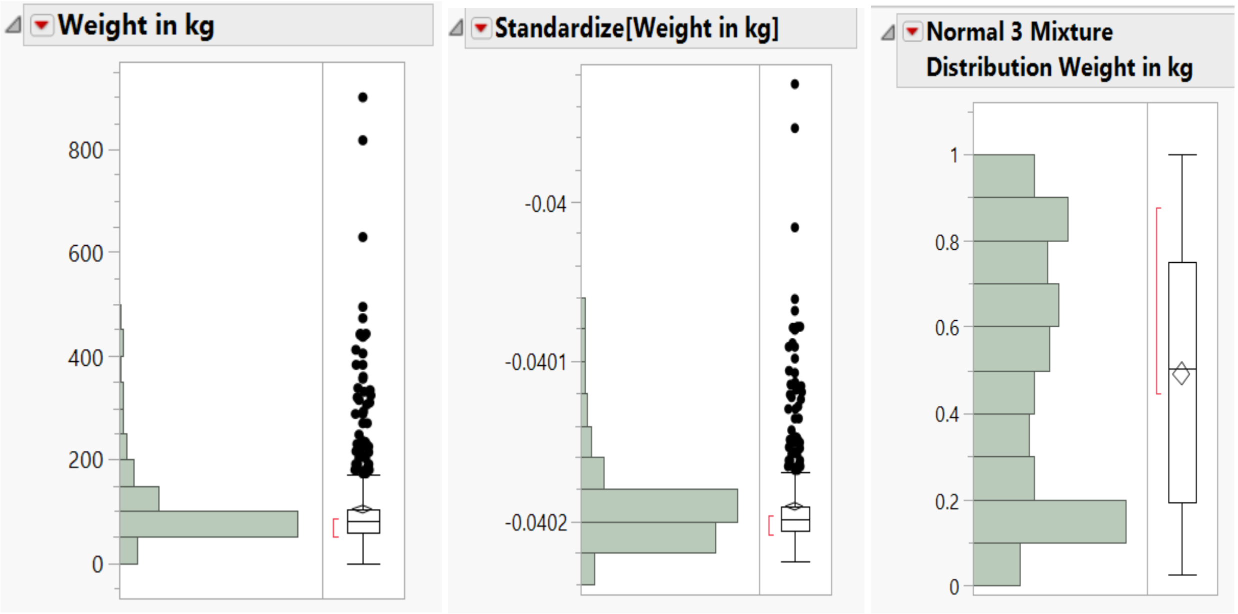
	Height in cm	Weight in kg
Height in cm	1.0000	0.5978
Weight in kg	0.5978	1.0000

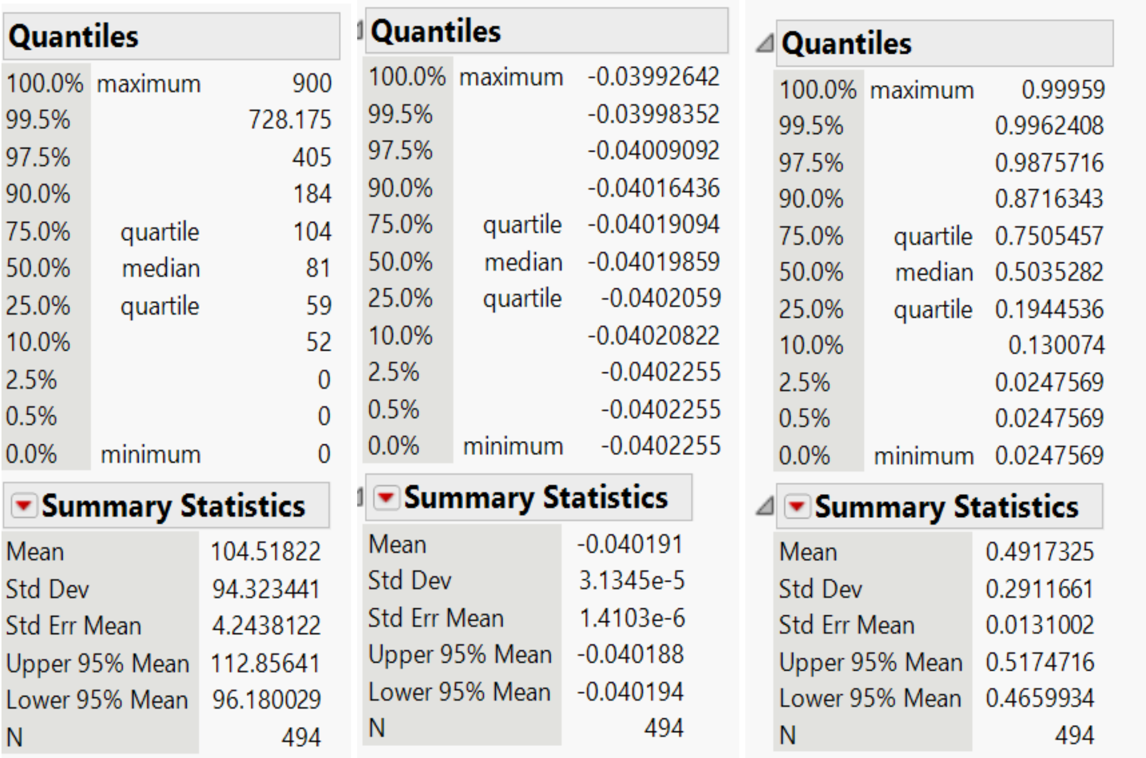
The correlations are estimated by Row-wise method.

Scatterplot Matrix

5. Standardize the Weight in kg column. Transform the Weight in kg column using the best fit that allows you to save the transform. Take a screen shot of the three distributions. Compare and contrast the three distributions. What has changed or stayed the same compared to the original column. What are the new means, medians, and standard deviations? How do outliers look across the three distributions?

“The work contained and presented here is my work and my work alone.”





Quantiles

100.0%	maximum	900
99.5%		728.175
97.5%		405
90.0%		184
75.0%	quartile	104
50.0%	median	81
25.0%	quartile	59
10.0%		52
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	104.51822
Std Dev	94.323441
Std Err Mean	4.2438122
Upper 95% Mean	112.85641
Lower 95% Mean	96.180029
N	494

Quantiles

100.0%	maximum	-0.03992642
99.5%		-0.03998352
97.5%		-0.04009092
90.0%		-0.04016436
75.0%	quartile	-0.04019094
50.0%	median	-0.04019859
25.0%	quartile	-0.0402059
10.0%		-0.04020822
2.5%		-0.0402255
0.5%		-0.0402255
0.0%	minimum	-0.0402255

Summary Statistics

Mean	-0.040191
Std Dev	3.1345e-5
Std Err Mean	1.4103e-6
Upper 95% Mean	-0.040188
Lower 95% Mean	-0.040194
N	494

Quantiles

100.0%	maximum	0.99959
99.5%		0.9962408
97.5%		0.9875716
90.0%		0.8716343
75.0%	quartile	0.7505457
50.0%	median	0.5035282
25.0%	quartile	0.1944536
10.0%		0.130074
2.5%		0.0247569
0.5%		0.0247569
0.0%	minimum	0.0247569

Summary Statistics

Mean	0.4917325
Std Dev	0.2911661
Std Err Mean	0.0131002
Upper 95% Mean	0.5174716
Lower 95% Mean	0.4659934
N	494

Sol:- Analyzing all the three distributions in box plot, there’s not much bigger difference in the weight in kg and normalized weight in kg, but there is a difference after using the best fit, because the values have been uniformly distributed in a bell shape. And the outliers have been included in the box after using Normal 3 mixture, comparing the mean, median and standard deviation there is a great difference in the values, the new mean is 0.4917. median in 0.503 and std dev is 0.29. comparatively the values have been lowered and are fitted in the range of 0 and 1. And the outliers were cutdown after using the best fit.

“The work contained and presented here is my work and my work alone.”

6. Eye color has 23 possible values in the non-excluded rows. If we were using this column to predict Alignment, how would you reduce the complexity in this column? What categories would you combine into what new categories? Justify your choices with screen shots of your work.

Sol:- The characters having with below colors are combined and categorized as Good Alignment characters.

Blue/White Alignment is good
Brown
Green/blue/
Hazel
Silver
Yellow/blue

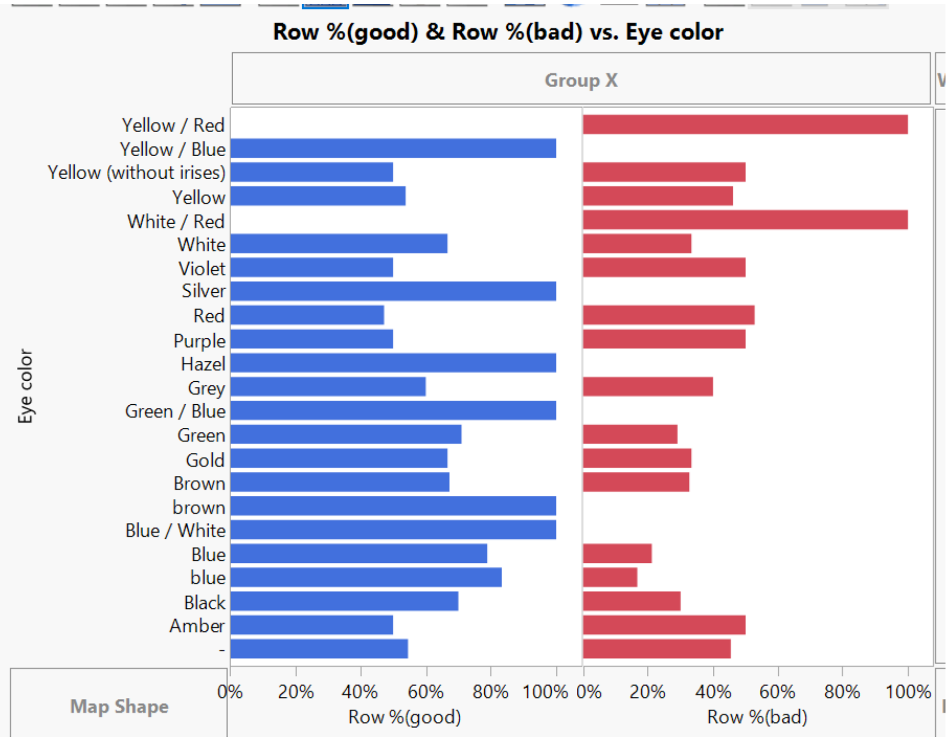
The characters having with below colors are combined and categorized as Bad Alignment characters.

White/red
Yellow/red

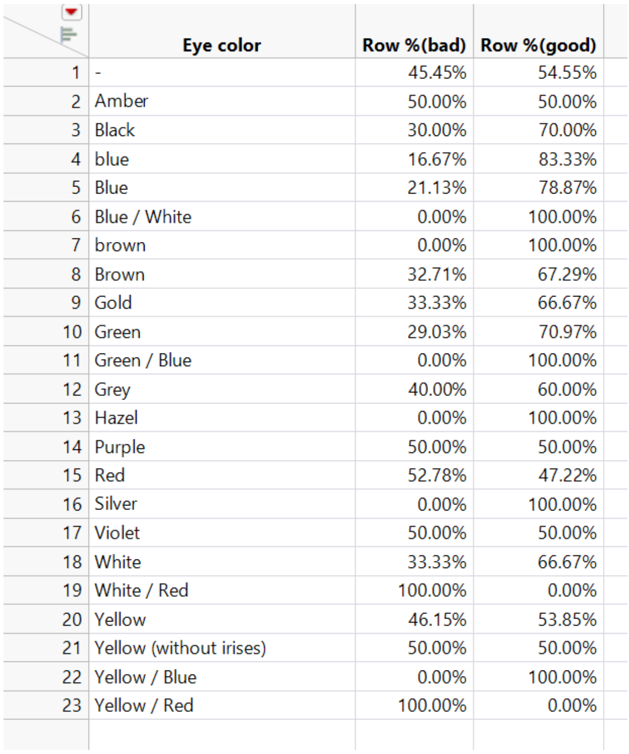
The characters having with below colors exactly share a equal share of 50% are combined and categorized as 50% Good & 50% Bad alignment characters.

Yellow (without Iris
Violet
Purple
amber

The total number of dummy variables could be N-1(1.e, 23-1 =22 dummy variables) can be made out, of which we have categorized 12 colors into 3 categories and the rest of the colors are 10, so we have reduced the complexity from 22 categories to 13 categories.

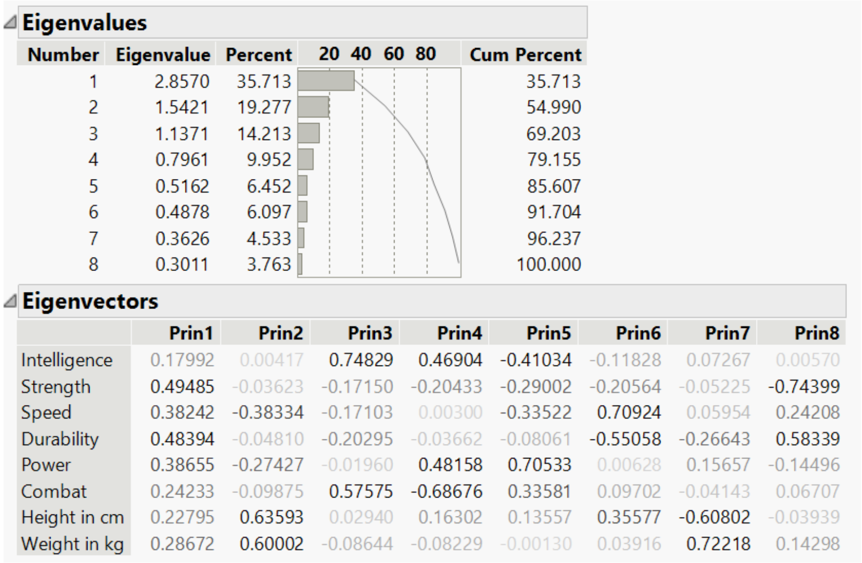


“The work contained and presented here is my work and my work alone.”



	Eye color	Row %(bad)	Row %(good)
1	-	45.45%	54.55%
2	Amber	50.00%	50.00%
3	Black	30.00%	70.00%
4	blue	16.67%	83.33%
5	Blue	21.13%	78.87%
6	Blue / White	0.00%	100.00%
7	brown	0.00%	100.00%
8	Brown	32.71%	67.29%
9	Gold	33.33%	66.67%
10	Green	29.03%	70.97%
11	Green / Blue	0.00%	100.00%
12	Grey	40.00%	60.00%
13	Hazel	0.00%	100.00%
14	Purple	50.00%	50.00%
15	Red	52.78%	47.22%
16	Silver	0.00%	100.00%
17	Violet	50.00%	50.00%
18	White	33.33%	66.67%
19	White / Red	100.00%	0.00%
20	Yellow	46.15%	53.85%
21	Yellow (without irises)	50.00%	50.00%
22	Yellow / Blue	0.00%	100.00%
23	Yellow / Red	100.00%	0.00%

7. Conduct a principal components analysis using Intelligence, Strength, Speed, Durability, Power, Combat, Height in cm, and Weight in kg. Comment on the results. Include a screen shot of the Eigenvectors and Eigenvalues (with cumulative percents). How much variability can you maintain if you keep 6 principal components? What variables contribute the most information to PC1 and PC2?



Eigenvalues

Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	2.8570	35.713		35.713
2	1.5421	19.277		54.990
3	1.1371	14.213		69.203
4	0.7961	9.952		79.155
5	0.5162	6.452		85.607
6	0.4878	6.097		91.704
7	0.3626	4.533		96.237
8	0.3011	3.763		100.000

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8
Intelligence	0.17992	0.00417	0.74829	0.46904	-0.41034	-0.11828	0.07267	0.00570
Strength	0.49485	-0.03623	-0.17150	-0.20433	-0.29002	-0.20564	-0.05225	-0.74399
Speed	0.38242	-0.38334	-0.17103	0.00300	-0.33522	0.70924	0.05954	0.24208
Durability	0.48394	-0.04810	-0.20295	-0.03662	-0.08061	-0.55058	-0.26643	0.58339
Power	0.38655	-0.27427	-0.01960	0.48158	0.70533	0.00628	0.15657	-0.14496
Combat	0.24233	-0.09875	0.57575	-0.68676	0.33581	0.09702	-0.04143	0.06707
Height in cm	0.22795	0.63593	0.02940	0.16302	0.13557	0.35577	-0.60802	-0.03939
Weight in kg	0.28672	0.60002	-0.08644	-0.08229	-0.00130	0.03916	0.72218	0.14298

Sol:- on observing the results, we can attain a cumulative percent by 91.70 even considering only 6 eigenvalues. The strength in 8th principal component being negative value of -0.743 has a stronger correlation and contribute the most since the farthest from zero in either direction are strongly correlated the first 5 principal components account for over 85%. If I keep 6 principal components, I can restore

“The work contained and presented here is my work and my work alone.”

91.704% of data, and loss of 8.296%. The key variables that contribute most to PC1 and PC2 are Strength, durability and Height in cm, weight in /kg respectively. The eigenvector with the largest eigenvalue is the direction with most variability.

8. Imagine a dataset with 30,000 rows. The target variable has 3,000 Yes and 27,000 No. If you were to partition the data with a 60/20/20 spilt and 50/50 oversampling in Training and Validation, how many No and Yes records would be in each partition?

Sol:- T**raining set**

Yes	1800
No	1800

Validation Set

Yes	600
No	600

Test Set

Yes	600
No	5400

“The work contained and presented here is my work and my work alone.”

Vinay Reddy Vangala

NetId: vrv22001