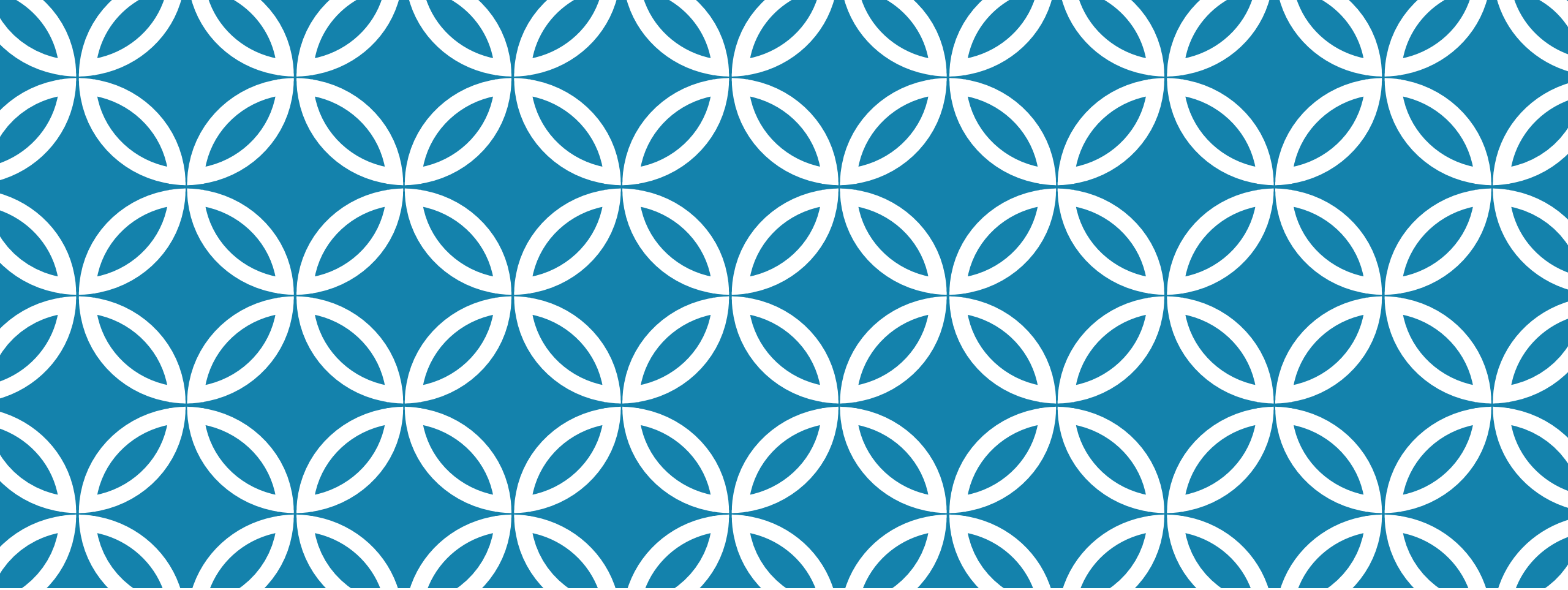


CORRELATION ANALYSIS

(DESCRIPTIVE STATISTIC ANALYSIS)

Maryam Najimigoshtasb

Fitness data

Various measures of heart and pulse rate were taken on men in a physical fitness course at N.C. State Univ.

Runpulse–Pulse rate while running
maxpulse–Maximum pulse rate
rstpulse–Resting pulse rate
oxy –oxygen consumption

We want to check the correlation between these measure and other attributes of men.

The data is as shown here.

Before checking the correlation we need go through following **steps.**

1. Check the causation
2. Check the limitation of the Pearson correlation
3. Make the hypothesis and apply the proper correlation method.

Obs	age	weight	oxy	runtime	rstpulse	runpulse	maxpulse	case
1	44	89.47	44.609	11.37	62	178	182	1
2	40	75.07	45.313	10.07	62	185	185	2
3	44	85.84	54.297	8.65	45	156	168	3
4	42	68.15	59.571	8.17	40	166	172	4
5	38	89.02	49.874	9.22	55	178	180	5
6	47	77.45	44.811	11.63	58	176	176	6
7	40	75.98	45.681	11.95	70	176	180	7
8	43	81.19	49.091	10.85	64	162	170	8
9	44	81.42	39.442	13.08	63	174	176	9
10	38	81.87	60.055	8.63	48	170	186	10
11	44	73.03	50.541	10.13	45	168	168	11
12	45	87.66	37.388	14.03	56	186	192	12
13	45	66.45	44.754	11.12	51	176	176	13
14	47	79.15	47.273	10.60	47	162	164	14
15	54	83.12	51.855	10.33	50	166	170	15
16	49	81.42	49.156	8.95	44	180	185	16
17	51	69.63	40.836	10.95	57	168	172	17
18	51	77.91	46.672	10.00	48	162	168	18
19	48	91.63	46.774	10.25	48	162	164	19
20	49	73.37	50.388	10.08	67	168	168	20
21	57	73.37	39.407	12.63	58	174	176	21
22	54	79.38	46.080	11.17	62	156	165	22
23	52	76.32	45.441	9.63	48	164	166	23
24	50	70.87	54.625	8.92	48	146	155	24
25	51	67.25	45.118	11.08	48	172	172	25
26	54	91.63	39.203	12.88	44	168	172	26
27	51	73.71	45.790	10.47	59	186	188	27
28	57	59.08	50.545	9.93	49	148	155	28
29	49	76.32	48.673	9.40	56	186	188	29
30	48	61.24	47.920	11.50	52	170	176	30
31	52	82.78	47.467	10.50	53	170	172	31

Causation

There are 21 combinations of variables, we need to look for any logical relationship between each two variables, If we could define and explain **any logical relationship** between two variable then we can apply correlation test. We will explain the causation of some here.

Age-weight: age impacts on weight as for younger people metabolism is higher than elderly and functionality of person is different as well. Therefore, age affects on weight. But it is not always true. There is causation but this causation is different person by person. Therefore we can measure the associations.

Age-Oxy: age affect on organs so most of the old people have heart and lung problem which impact on oxygen consumptions. Therefore we can measure the associations.

weight- oxy: higher weight hard breathing and low oxy consumption
Therefore we can measure the associations.

Oxy-runtime: long time running more heart pumping an more oxygen needed but person are able to take less oxygen Therefore we can measure the associations.

Assumptions or limitations of Pearson's correlations :

1. Having sufficient sample size

2. There should not be any outliers in the data or attributes

3. There should not be any non-linear relationship between each variables

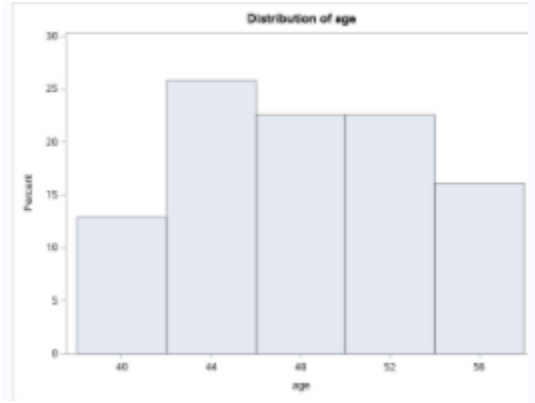
First Assumption:

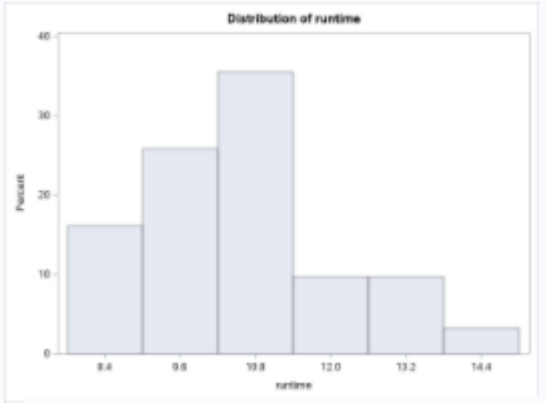
However, the sample size is not large. it is adequate to analyze the correlation test. We need to be careful about the reliability of the correlation result.

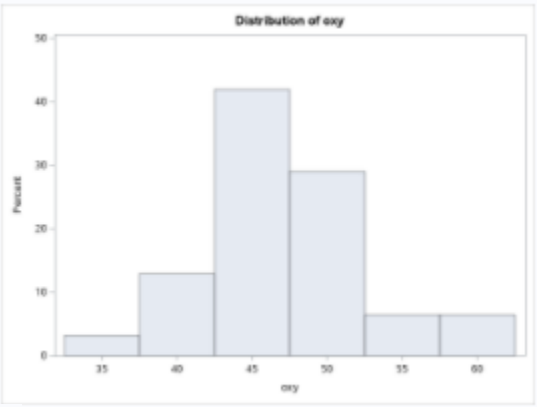
Obs	age	weight	oxy	runtime	rstpulse	runpulse	maxpulse	case
1	44	89.47	44.609	11.37	62	178	182	1
2	40	75.07	45.313	10.07	62	185	185	2
3	44	85.84	54.297	8.65	45	156	168	3
4	42	68.15	59.571	8.17	40	166	172	4
5	38	89.02	49.874	9.22	55	178	180	5
6	47	77.45	44.811	11.63	58	176	176	6
7	40	75.98	45.681	11.95	70	176	180	7
8	43	81.19	49.091	10.85	64	162	170	8
9	44	81.42	39.442	13.08	63	174	176	9
10	38	81.87	60.055	8.63	48	170	186	10
11	44	73.03	50.541	10.13	45	168	168	11
12	45	87.66	37.388	14.03	56	186	192	12
13	45	66.45	44.754	11.12	51	176	176	13
14	47	79.15	47.273	10.60	47	162	164	14
15	54	83.12	51.855	10.33	50	166	170	15
16	49	81.42	49.156	8.95	44	180	185	16
17	51	69.63	40.836	10.95	57	168	172	17
18	51	77.91	46.672	10.00	48	162	168	18
19	48	91.63	46.774	10.25	48	162	164	19
20	49	73.37	50.388	10.08	67	168	168	20
21	57	73.37	39.407	12.63	58	174	176	21
22	54	79.38	46.080	11.17	62	156	165	22
23	52	76.32	45.441	9.63	48	164	166	23
24	50	70.87	54.625	8.92	48	146	155	24
25	51	67.25	45.118	11.08	48	172	172	25
26	54	91.63	39.203	12.88	44	168	172	26
27	51	73.71	45.790	10.47	59	186	188	27
28	57	59.08	50.545	9.93	49	148	155	28
29	49	76.32	48.673	9.40	56	186	188	29
30	48	61.24	47.920	11.50	52	170	176	30
31	52	82.78	47.467	10.50	53	170	172	31

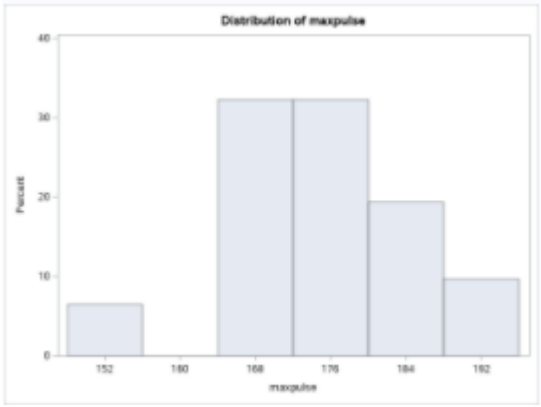
Second Assumption:

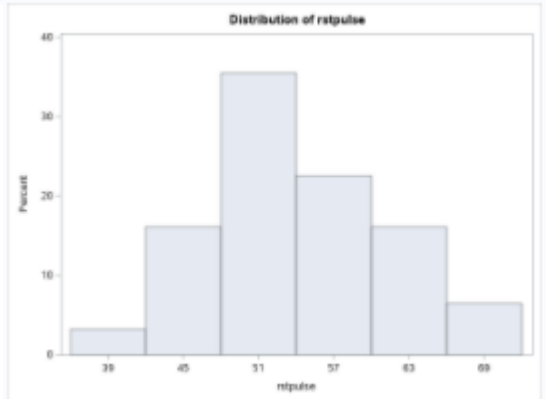
As we can see here there is not outliers in the data.
To check the outliers we can also check other statistics checking the extreme observations , the range and standard deviation and so on.

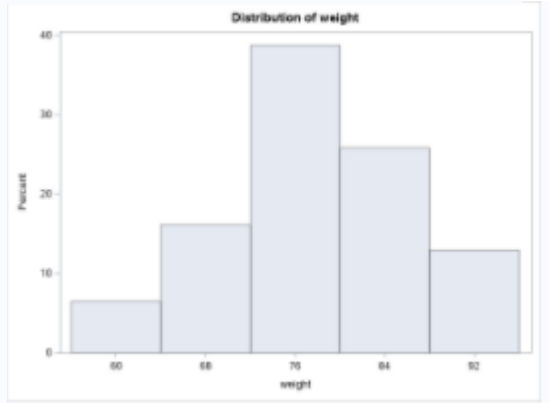


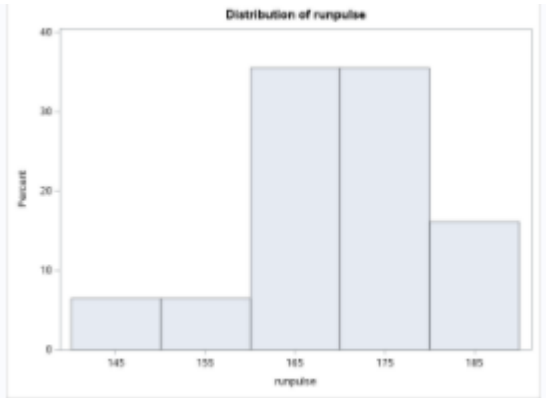






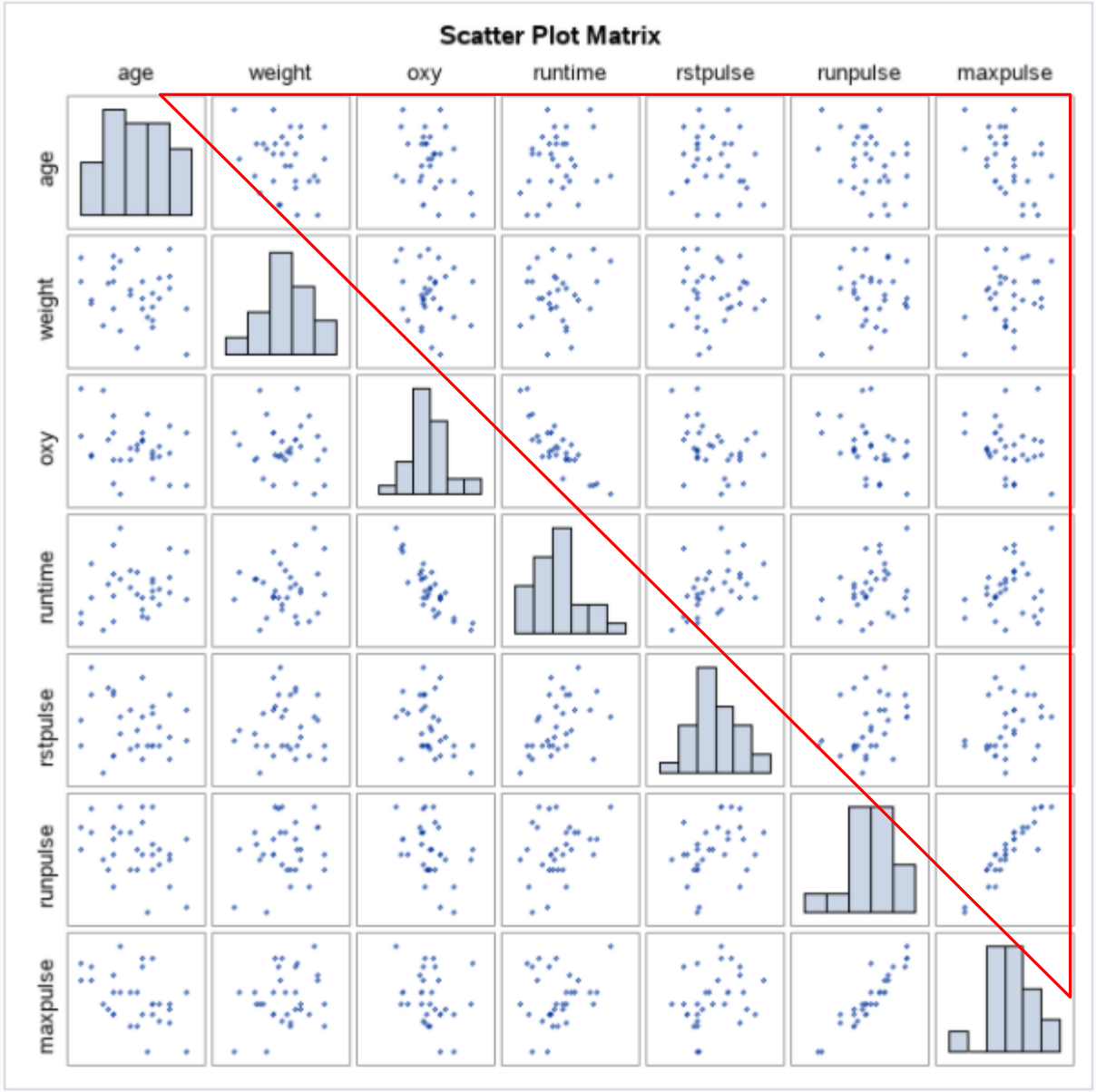






Third Assumption:

There is no any non-linear pattern in each plot .
The clear patterns are linear which are between Oxy and runtime, runpulse and maxpulse.



Make the Hypothesis

- ✓ H0:(Rho) $\rho = 0$ ("the population correlation coefficient is 0; there is no association")
- ✓ H1: (Rho) $\rho \neq 0$ ("the population correlation coefficient is not 0; a nonzero correlation could exist")

Where ρ is the population correlation coefficient.

- we should have 21 hypothesis for each pairwise. As we have 7 variables.
- Considering the level of significant of 0.05.

The result of the Pearson correlation.

Strong correlation are colored by green :

runtime and oxy : the correlation between **runtime and oxy** is very strong and they are in the oppose direction. it means if runtime increase the oxy will decrease which make sense as we know there is a causation which indicates when runtime increase the hearth pump faster, and the consumption of oxygen will reduce. the probability of accepting the hypothesis id lower than the level of significant so **we reject the H0.**

Maxpulse and runpulse: the correlation between **Maxpulse and runpulse** is very strong and they are in the same(positive) direction. it means if Maxpulse increase the runpulse will increase which make sense as we know there is a causation which indicates when Maximum pulse rate increase the Pulse rate while running will increase. the probability of accepting the hypothesis id lower than the level of significant so **we reject the H0.**

Pearson Correlation Coefficients, N = 31
Prob > |r| under H0: Rho=0

	age	weight	oxy	runtime	rstpulse	runpulse	maxpulse
age	1.00000	-0.23354 0.2061	-0.30459 0.0957	0.18875 0.3092	-0.16410 0.3777	-0.33787 0.0630	-0.43292 0.0150
weight	-0.23354 0.2061	1.00000	-0.16275 0.3817	0.14351 0.4412	0.04397 0.8143	0.18152 0.3284	0.24938 0.1761
oxy	-0.30459 0.0957	-0.16275 0.3817	1.00000	-0.86219 <.0001	-0.39936 0.0260	-0.39797 0.0266	-0.23674 0.1997
runtime	0.18875 0.3092	0.14351 0.4412	-0.86219 <.0001	1.00000	0.45038 0.0110	0.31365 0.0858	0.22610 0.2213
rstpulse	-0.16410 0.3777	0.04397 0.8143	-0.39936 0.0260	0.45038 0.0110	1.00000	0.35246 0.0518	0.30512 0.0951
runpulse	-0.33787 0.0630	0.18152 0.3284	-0.39797 0.0266	0.31365 0.0858	0.35246 0.0518	1.00000	0.92975 <.0001
maxpulse	-0.43292 0.0150	0.24938 0.1761	-0.23674 0.1997	0.22610 0.2213	0.30512 0.0951	0.92975 <.0001	1.00000

Weak correlation colored by light pink and yellow:

Light pink: the probability of the accepting hypothesis is higher than level of significant so we **fail to reject the hypothesis** so we can not interpret the correlation figure. **In fact there is dependency between variables**

Yellow: the probability of the accepting hypothesis is lower than level of significant so we **reject the hypothesis** so we can say there is weak correlation between variables respecting to the the −/+sign which indicate positive or negative correlation

Pearson Correlation Coefficients, N = 31
Prob > |r| under H0: Rho=0

	age	weight	oxy	runtime	rstpulse	runpulse	maxpulse
age	1.00000	-0.23354 0.2061	-0.30459 0.0957	0.18875 0.3092	-0.16410 0.3777	-0.33787 0.0630	-0.43292 0.0150
weight	-0.23354 0.2061	1.00000	-0.16275 0.3817	0.14351 0.4412	0.04397 0.8143	0.18152 0.3284	0.24938 0.1761
oxy	-0.30459 0.0957	-0.16275 0.3817	1.00000	-0.86219 <.0001	-0.39936 0.0260	-0.39797 0.0266	-0.23674 0.1997
runtime	0.18875 0.3092	0.14351 0.4412	-0.86219 <.0001	1.00000	0.45038 0.0110	0.31365 0.0858	0.22610 0.2213
rstpulse	-0.16410 0.3777	0.04397 0.8143	-0.39936 0.0260	0.45038 0.0110	1.00000	0.35246 0.0518	0.30512 0.0951
runpulse	-0.33787 0.0630	0.18152 0.3284	-0.39797 0.0266	0.31365 0.0858	0.35246 0.0518	1.00000	0.92975 <.0001
maxpulse	-0.43292 0.0150	0.24938 0.1761	-0.23674 0.1997	0.22610 0.2213	0.30512 0.0951	0.92975 <.0001	1.00000

Very weak correlation **dark pink**

dark pink: the probability of the accepting hypothesis is higher than level of significant so we **fail to reject the hypothesis** so we can not interpret the correlation figure. In fact there is dependency between variables

Pearson Correlation Coefficients, N = 31
Prob > |r| under H0: Rho=0

	age	weight	oxy	runtime	rstpulse	runpulse	maxpulse
age	1.00000	-0.23354 0.2061	-0.30459 0.0957	0.18875 0.3092	-0.16410 0.3777	-0.33787 0.0630	-0.43292 0.0150
weight	-0.23354 0.2061	1.00000	-0.16275 0.3817	0.14351 0.4412	0.04397 0.8143	0.18152 0.3284	0.24938 0.1761
oxy	-0.30459 0.0957	-0.16275 0.3817	1.00000	-0.86219 <.0001	-0.39936 0.0260	-0.39797 0.0266	-0.23674 0.1997
runtime	0.18875 0.3092	0.14351 0.4412	-0.86219 <.0001	1.00000	0.45038 0.0110	0.31365 0.0858	0.22610 0.2213
rstpulse	-0.16410 0.3777	0.04397 0.8143	-0.39936 0.0260	0.45038 0.0110	1.00000	0.35246 0.0518	0.30512 0.0951
runpulse	-0.33787 0.0630	0.18152 0.3284	-0.39797 0.0266	0.31365 0.0858	0.35246 0.0518	1.00000	0.92975 <.0001
maxpulse	-0.43292 0.0150	0.24938 0.1761	-0.23674 0.1997	0.22610 0.2213	0.30512 0.0951	0.92975 <.0001	1.00000

Best Regards

Maryam Najimigoshtasb.