Metadata/ Data Dictionary

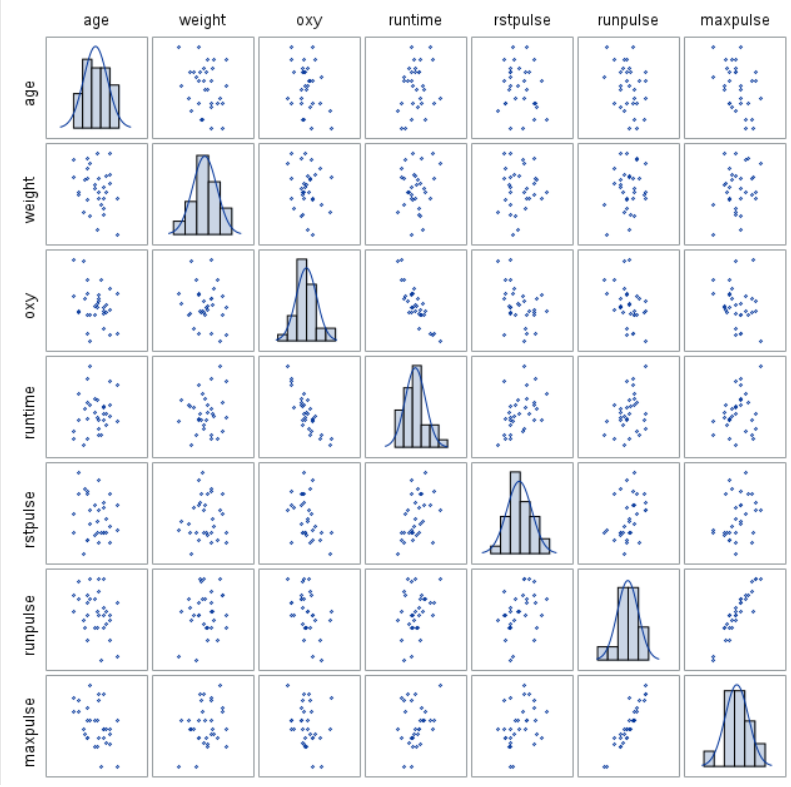
Variable	Description
Age	Age in years
Weight	Weight in kg
Oxy	Oxygen consumption
Runtime	Running time in minutes and seconds
Rstpulse	Resting pulse rate
Runpulse	Pulse rate while running
Maxpulse	Maximum pulse rate

1. Identifying relationships between variables

The goal of the assignment is to identify if there are relationships between variables, and determine how strong the relationships are (if any). The first step is to generate a scatter plot matrix to identify if there is linearity within the data, and choose between the Pearson or Spearman coefficient.

SAS Code:

```
proc corr data = fitness pearson

var age weight oxy runtime rstpulse runpulse maxpulse;

run;

proc sgscatter data=fitness;

matrix age weight oxy runtime rstpulse runpulse maxpulse / diagonal=(histogram normal);

run;
```



Looking at the plots in the scatterplot matrix, we can see that the data looks linear and there is no L shaped curve in the matrix, thus we choose to use Pearson's correlation instead.

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

Hypothesis Testing:

Looking at the coefficient value and the p-values for the columns and rows, we must use a hypothesis test first to determine if the results are statistically significant.

H0 Null hypothesis: Variable 1 is not associated with Variable 2

H1 Alternative hypothesis: Variable 1 is associated with Variable 2

Using a significance level of 0.05, as it is industry standard.

Groups with p-value below 0.05:

We can see that there are a few variable groups that have a p value below 0.05, and thus are statistically significant. We reject the null hypothesis for these groups, and determine that there is a linear association between the two variables.

These are as follows:

Age and Maxpulse: 0.0150 p-value

Oxy and runtime: <0.0001 p-value

Oxy and rstpulse: 0.0260 p-value

Oxy and runpulse: 0.0266 p-value

Runtime and rstpulse: 0.0110 p-value

Runpulse and maxpulse: <0.001 p-value

The other groups have a p-value larger than 0.05, thus are not statistically significant. The correlation coefficient for those variable groups do not provide a reliable method of predicting correlation between the variables.

Age and Maxpulse:

H0 Null hypothesis: Age is not associated with Maxpulse

H1 Alternative hypothesis: Age is associated with Maxpulse

As the p-value is 0.0150 and less than the significance level of 0.05, we reject the null hypothesis and identify that age is associated with maxpulse.

The Pearson correlation coefficient of -0.43292 shows that there is a weak negative correlation between age and maxpulse, such that when age increases, the maxpulse of an individual decreases slightly. The interpretation of this result could be due to less elastic heart muscles as an individual ages, reducing the pulse rate of older individuals.

Oxy and runtime:

H0 Null hypothesis: Oxygen consumption is not associated with the running time

H1 Alternative hypothesis: Oxygen consumption is associated with the running time

As the p-value is <0.0001 and less than the significance level of 0.05, we reject the null hypothesis and identify that oxygen consumption is associated with the running time.

The Pearson correlation coefficient of -0.86219 shows that there is a strong negative correlation between oxygen consumption and running time, showing that the longer an individual runs, they will consume less oxygen at a much lower consumption level. This can be explained by individuals being more physically fit when they can run for longer periods of time, such that the efficiency and capacity of their lungs is higher, and the oxygen consumption in their day to day activities is less.

Oxy and Rstpulse:

H0 Null hypothesis: Oxygen consumption is not associated with the resting pulse rate

H1 Alternative hypothesis: Oxygen consumption is associated with the resting pulse rate

As the p-value is 0.0260 and less than the significance level of 0.05, we reject the null hypothesis and identify that oxygen consumption is associated with the resting pulse rate.

The Pearson correlation coefficient of -0.39936 shows that there is a weak negative correlation between oxygen consumption and resting pulse rate, showing that individuals with higher oxygen consumption tend to have slightly less resting pulse rates. This could be explained by individuals with lower resting pulse rate being more physically fit, thus the efficiency and capacity of their lungs is higher, and the oxygen consumption in their day to day activities is less.

Oxy and Runpulse:

H0 Null hypothesis: Oxygen consumption is not associated with the running pulse rate

H1 Alternative hypothesis: Oxygen consumption is associated with the running pulse rate

As the p-value is 0.0266 and less than the significance level of 0.05, we reject the null hypothesis and identify that oxygen consumption is associated with the running pulse rate.

The Pearson correlation coefficient of -0.39797 shows that there is a weak negative correlation between oxygen consumption and running pulse rate, showing that individuals with higher oxygen consumption tend to have slightly less running pulse rates. Similar to the resting pulse rate explanation, individuals with a lower running pulse rate tend to be more physically fit, thus the efficiency and capacity of their lungs is higher, and the oxygen consumption in their day to day activities is less.

Runtime and Rstpulse:

H0 Null hypothesis: Running time is not associated with the resting pulse rate

H1 Alternative hypothesis: Running time is associated with the resting pulse rate

As the p-value is 0.0110 and less than the significance level of 0.05, we reject the null hypothesis and identify that running time is associated with the resting pulse rate.

The Pearson correlation coefficient of 0.45038 shows that there is a weak positive correlation between running time and resting pulse rate, showing that individuals with higher running time tend to have slightly higher resting pulse rates. Generally, a lower resting pulse rate is an indicator of being more physically fit and can result in a higher running time, however as the correlation is weak positive, it may not be the sole determinant of a higher running time.

Runpulse and Maxpulse:

H0 Null hypothesis: Running pulse rate is not associated with the maximum pulse rate

H1 Alternative hypothesis: Running pulse rate is associated with the maximum pulse rate

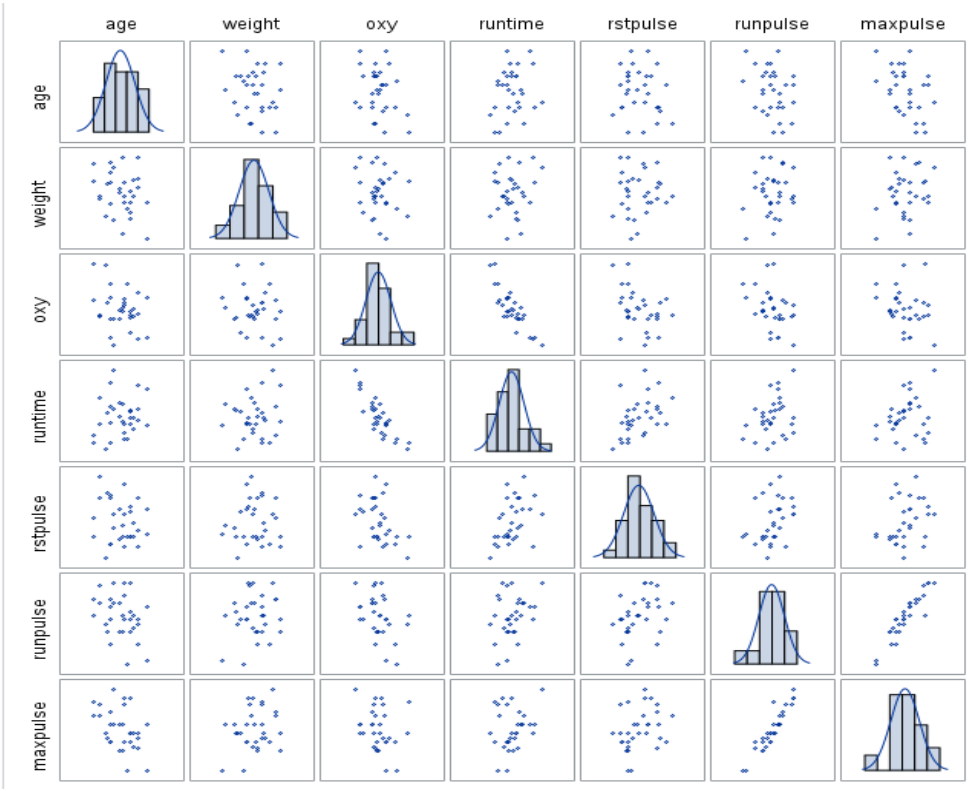
As the p-value is <0.0001 and less than the significance level of 0.05, we reject the null hypothesis and identify that running pulse rate is associated with the maximum pulse rate.

The Pearson correlation coefficient of 0.92975 shows that there is a strong positive correlation between running pulse rate and maximum pulse rate, showing that individuals with higher running pulse rate tend to have much higher maximum pulse rates. This is because running increases the

pulse rate due to higher oxygen demand, and a higher pulse rate will be reflected in the maximum pulse rate the individual can achieve.

Checking the limitations of correlation coefficient:

Using Pearson's correlation coefficient comes with three limitations:



1. **Linearity: r is a measure of linear association**
    - There is a limitation that the strength of non-linear relationship cannot be truly determined by the Pearson correlation coefficient.
    - However, as seen above in the scatterplot matrix, the different scatterplots between variables show linearity. This is because the data points mainly follow a straight line or cloud shape, rather than an L shaped curve. Thus, it can be determined that there is linearity between the groups of data.

2. **Sufficient sample size**
    - A small sample size can decrease the reliability of the r coefficient.
    - The sample size for each variable is 31, which is larger than 30 and thus this increases the reliability of the correlation coefficient. To further increase the reliability, a larger sample size could be obtained.

3. **Outliers**
    - Outliers can affect the r coefficient, decreasing the reliability of the coefficient result.
    - However, as seen in the scatterplot matrix, there are no outliers and the datapoints are matched quite closely with each other between the groups of variables. Thus this limitation can be overcome.

Summary of relationship between variables and their correlation:

Relationship between variables	**Correlation**
Age and Maximum pulse rate	Weak negative
Oxygen consumption and runtime	Strong negative
Oxygen consumption and resting pulse rate	Weak negative
Oxygen consumption and running pulse rate	Weak negative
Runtime and resting pulse rate	Weak positive
Running pulse rate and maximum pulse rate	Strong positive