Seneca

SENECA COLLEGE OF APPLIED ARTS AND TECHNOLOGY

SENECA BUSINESS

BAN100 - Statistics for Analytics Other

Version NA

DATE: 7/23/2023 TIME ALLOWED: 14 days

PROFESSOR(S): Samaneh Gholami

Allowable Examination Aids: (check applicable boxes)

☒ Calculators (non-programmable only) ☒ Math Tables (normal distribution table) ☒ Periodic Tables

☐ Formula Sheets (attached) ☒ Textbooks ☒ Probability Tables

☒ Dictionary ☒ Notes ☐ Other

Answers to be completed on:

☐ Exam Booklet ☐ GradeMaster Card ☐ Exam Paper

TOTAL MARKS: 20 WEIGHTED VALUE: 20

INSTRUCTIONS:

Academic Integrity Policy. Seneca upholds a learning community that values academic integrity, honesty, fairness, trust, respect, responsibility and courage. These values enhance Seneca's commitment to students by delivering high-quality education and teaching excellence, while supporting a positive learning environment. The AI policy is always in effect. Note Sections 2.3 and 2.4:

"…2.3 Should there be a suspected violation of this policy (e.g.…cheating, falsification, impersonation or plagiarism), the academic integrity sanctions will be applied according to the severity of the offence committed. Refer to Appendix B for the academic integrity sanctions. 2.4 Should a suspected violation of this policy be a result of, or in combination with, a suspected violation of Seneca's Student Code of Conduct and/or another non-academic-related Seneca policy, the matter will be investigated and adjudicated through the process found in the Student Code of Conduct."

TO BE COMPLETED BY STUDENT
SUBJECT SECTION NUMBER (e.g. QNM223 AA):
STUDENT NAME: Ugonna Okengwu
STUDENT NUMBER: 114939192
STUDENT SIGNATURE:

APPROVED BY: ____

Cristina Italia, Interim Chair
School of Management and Entrepreneurship

DATE: ____

Variables:

1. **Runpulse:** Pulse rate while running.
2. **Maxpulse**: Maximum pulse rate.
3. **Rstpulse**: Resting pulse rate.
4. **oxy**: oxygen consumption.

We will be describing the relationships between the variables. I was able to generate scatter plots and correlations for the variables, interpreting the results and checking the limitations of the correlation coefficient for the data.

Step 1: Hypothesis

We state the hypotheses for each variable pair.

H0 assumes no association ($r = 0$), implying that variable A is independent of variable B.

Ha suggests an association ($r \neq 0$), indicating that variable A is not independent of variable B, and co-dependency exists between them.

Step 2: Analysis:

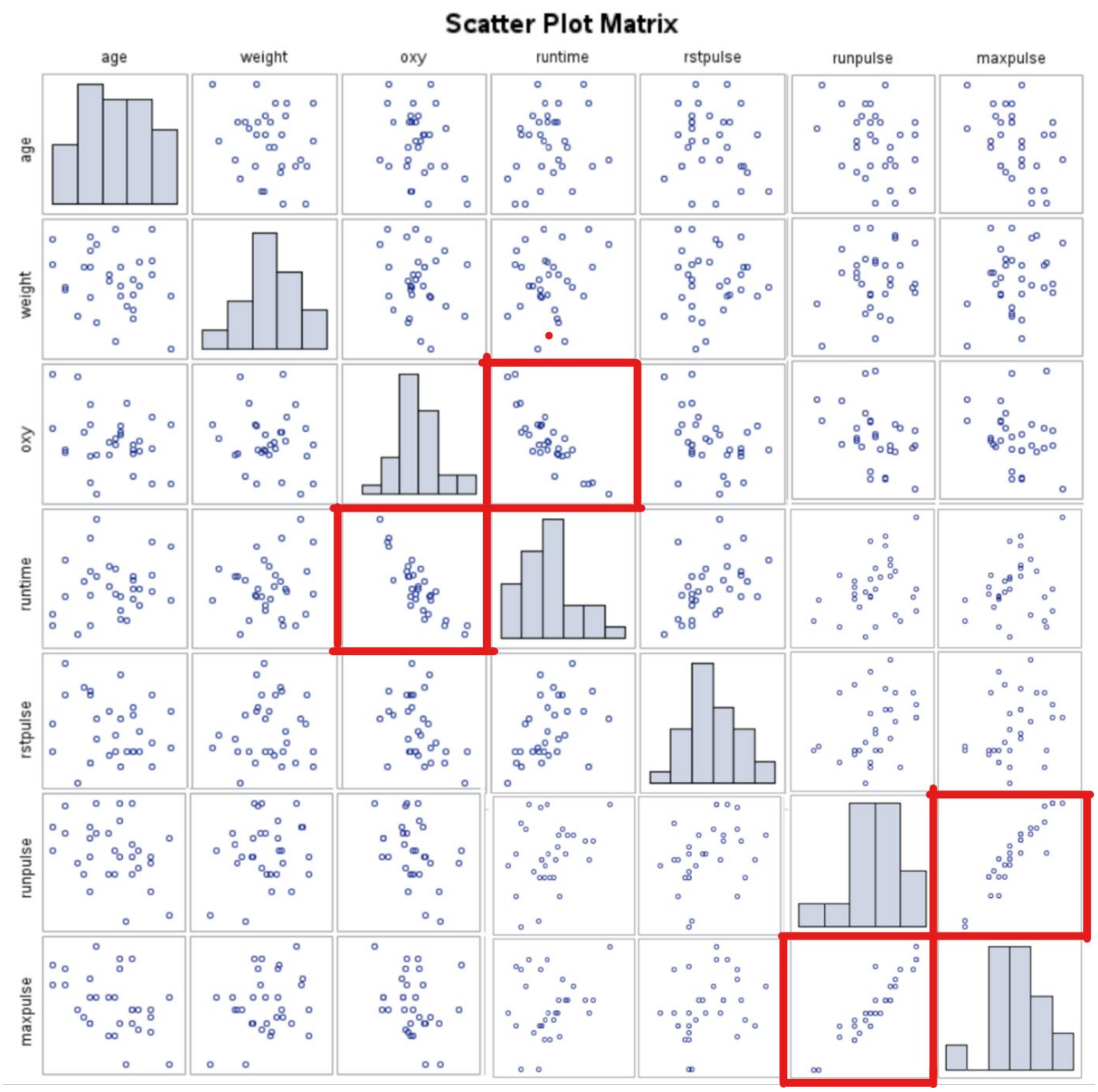
7/23/23, 12:47 PM Code: ASSIGNMENT3.sas

```
DATA fitness;
    input age weight oxy runtime rstpulse runpulse maxpulse;
    case = _n_;
    datalines;
    44 89.47 44.609 11.37 62 178 182
    40 75.07 45.313 10.07 62 185 185
    44 85.84 54.297 8.65 45 156 168
    42 68.15 59.571 8.17 40 166 172
    38 89.02 49.874 9.22 55 178 180
    47 77.45 44.811 11.63 58 176 176
    40 75.98 45.681 11.95 70 176 180
    43 81.19 49.091 10.85 64 162 170
    44 81.42 39.442 13.08 63 174 176
    38 81.87 60.055 8.63 48 170 186
    44 73.03 50.541 10.13 45 168 168
    45 87.66 37.388 14.03 56 186 192
    45 66.45 44.754 11.12 51 176 176
    47 79.15 47.273 10.60 47 162 164
    54 83.12 51.855 10.33 50 166 170
    49 81.42 49.156 8.95 44 180 185
    51 69.63 40.836 10.95 57 168 172
    51 77.91 46.672 10.00 48 162 168
    48 91.63 46.774 10.25 48 162 164
    49 73.37 50.388 10.08 67 168 168
    57 73.37 39.407 12.63 58 174 176
    54 79.38 46.080 11.17 62 156 165
    52 76.32 45.441 9.63 48 164 166
    50 70.87 54.625 8.92 48 146 155
    51 67.25 45.118 11.08 48 172 172
    54 91.63 39.203 12.88 44 168 172
    51 73.71 45.790 10.47 59 186 188
    57 59.08 50.545 9.93 49 148 155
    49 76.32 48.673 9.40 56 186 188
    48 61.24 47.920 11.50 52 170 176
    52 82.78 47.467 10.50 53 170 172
    ;
run;

proc corr data = Fitness plots = matrix(hist);
var age weight oxy runtime rstpulse runpulse maxpulse;
RUN;

PROC GPLOT DATA = Fitness;
plot runpulse*maxpulse;
run;

proc gplot data = Fitness;
plot runtime*oxy;
run;
```

about:blank 1/2

The analysis starts with visualization plots in SAS to examine linearity using scatter plots. Strong correlations are observed in certain variable pairs, specifically a positive correlation between run pulse and max pulse, and a negative correlation between runtime and oxygen consumption. Other variables show no clear trend.

Pearson correlation tables are generated to analyze and explain the results. Most values have no strong positive or negative correlation (r < 0.75), except for maxpulse-runpulse and oxygen-runtime, which exhibit significant associations.



As can be observed from the scatter plots above, only the boxes highlighted in red depict association or relatively strong correlation. We see that a clear trend or line is formed compared to the other variables. A positive correlation is formed between variables run pulse and max pulse. On the other hand, a negative trend or correlation can be observed between variables runtime and oxygen consumption. For other variables, a clear trend cannot be observed, just a cluster of groups and points.

The CORR Procedure

7 Variables:	age weight oxy runtime rstpulse runpulse maxpulse

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
age	31	47.67742	5.21144	1478	38.00000	57.00000
weight	31	77.44452	8.32857	2401	59.08000	91.63000
oxy	31	47.37581	5.32723	1469	37.38800	60.05500
runtime	31	10.58613	1.38741	328.17000	8.17000	14.03000
rstpulse	31	53.45161	7.61944	1657	40.00000	70.00000
runpulse	31	169.64516	10.25199	5259	146.00000	186.00000
maxpulse	31	173.77419	9.16410	5387	155.00000	192.00000

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

For each cell in the correlation matrix table above, we see two values. The upper one being r, the correlation coefficient and the lower one being the P value. Majority of the values are below 0.75 which means there is no strong positive or negative correlation among them. The only exceptions highlighted being maxpulse – runpulse and variables oxygen – runtime.

- **Association between maxpulse and runpulse:**

A strong positive correlation of 0.92975 indicates that as pulse rate while running increases, maximum pulse rate may also increase, particularly in fit individuals. They are more likely able to sustain a higher pulse rate while running. The reverse will be the case for people who are less fit physically.

The p-value ($P < 0.0001$) confirms significant variable association. Hence, we reject the null hypothesis that says there is no association between these 2 variables. Correlation coefficient of 0.92975 ≠ 0.

- **Association between oxygen consumption and run time:**

A strong negative correlation of -0.82619 suggests that extensive runtime leads to reduced oxygen consumption and vice versa. The p-value ($P < 0.0001$) indicates a significant association meaning the null hypothesis should be rejected. Association exists between the two variables with a measure of -0.82619. These correlation coefficients obtained can be deemed accurate since linearity and strong trends were achieved between these variables.

Other variable pairs have moderate, weak, or insignificant positive or negative correlations ($r \approx 0$). For example, there is no significant relationship between resting pulse rate and weight, despite the logical possibility of weight affecting pulse rate.

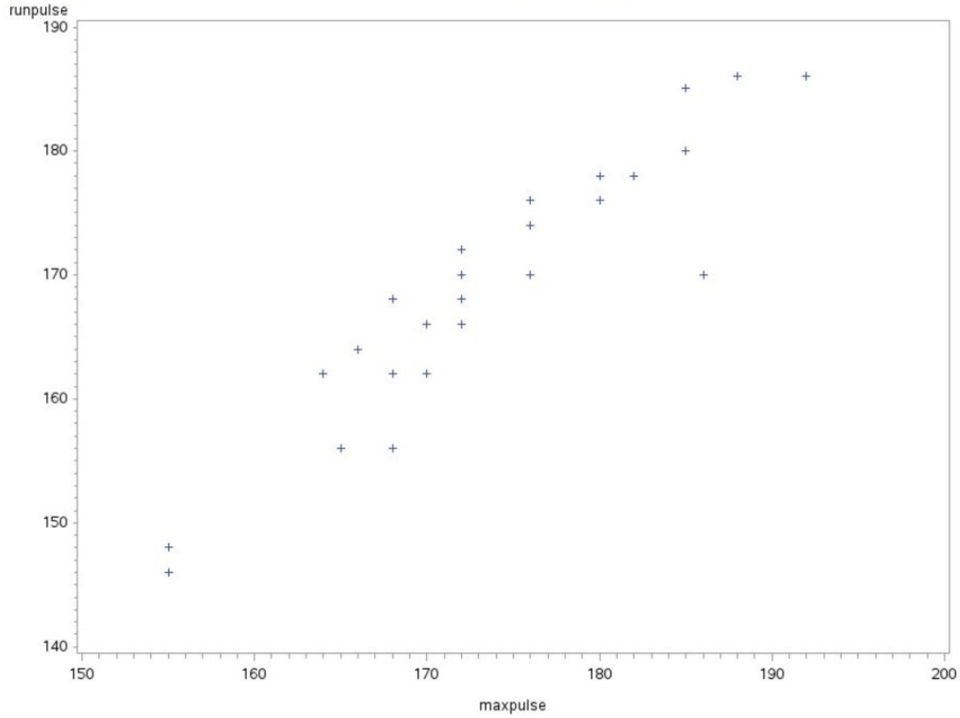
R = 0.04397. this is approximately 0 which shows no correlation.

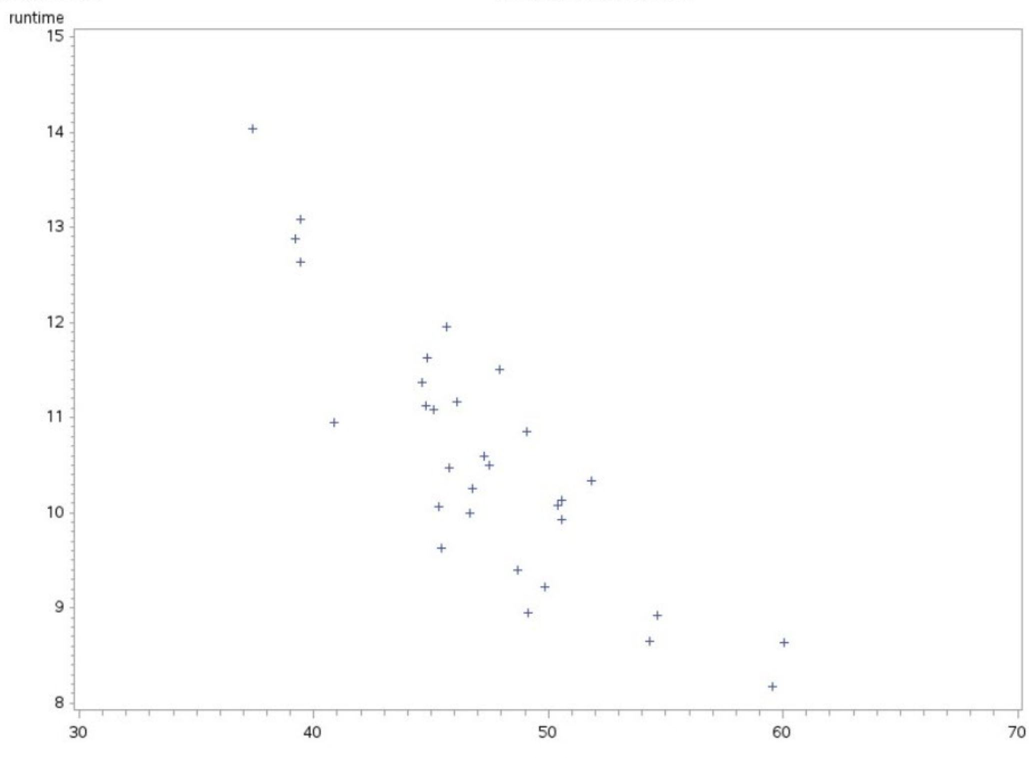
P= 0.8143 > 0.05 indicating that the alternative hypothesis should be rejected and r = 0

From a logical standpoint, there might be a relationship between a person's weight and pulse rate. In general, a higher body weight can increase the workload on the heart, leading to an increased resting pulse rate. This is because the heart must work harder to supply blood to the additional body mass. As a result, it might be possible to observe a positive correlation between weight and resting pulse rate.

Proc gplot substantiates the earlier strong correlations observed in the scatter plots matrix.

```
PROC GPLOT DATA = Fitness;
plot runpulse*maxpulse;
run;

proc gplot data = Fitness;
plot runtime*oxy;
run;
```


runpulse
190
180
170
160
150
140
150
160
170
180
190
200
maxpulse


runtime
15
14
13
12
11
10
9
8
30
40
50
60
70
oxy

Step 3: Further analysis using Spearman.

Due to nonlinearity observed in some variables, we run the Spearman correlation test to obtain more accurate results. Though the sample size is relatively sufficient, and no outliers are conspicuously visible, as the linearity condition was not satisfied, it is best to run the spearman test to confirm authentic results. The Spearman method is not limited by linearity.

The CORR Procedure

7 Variables:	age weight oxy runtime rstpulse runpulse maxpulse

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
age	31	47.67742	5.21144	1478	38.00000	57.00000
weight	31	77.44452	8.32857	2401	59.08000	91.63000
oxy	31	47.37581	5.32723	1469	37.38800	60.05500
runtime	31	10.58613	1.38741	328.17000	8.17000	14.03000
rstpulse	31	53.45161	7.61944	1657	40.00000	70.00000
runpulse	31	169.64516	10.25199	5259	146.00000	186.00000
maxpulse	31	173.77419	9.16410	5387	155.00000	192.00000

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

Above are the results obtained from the Spearman's correlation test. It can be observed that like the Pearson results, only 2 variable pairs show strong trend with Rho between 0.75 and 1. Spearman's correlation results largely confirm the Pearson results, showing strong associations in the same variable pairs.

Step 4: Interpretation of results

The results are summarized into three categories:

Insignificant association:

Variables with no significant relationship (r or ρ between 0 and 0.25). Examples include age-weight, age-oxygen, and weight-oxygen.

• Age and weight – negative insignificant association. There is no association between these two variables. Rho is approximately 0.
• Age and oxygen – negative insignificant association. Rho ≈ 0 (-16% correlation)
• Age and Runtime – Positive insignificant association. Rho ≈ 0 with a 15% strength of association
• Age and restpulse – negative insignificant association. Rho ≈ 0 with a 12% likelihood of association
• Weight and oxygen – negative insignificant association. Rho ≈ 0 with a 9% strength of association. Increase in weight resulting in lower oxygen consumption. However, this is more unlikely than likely as there is extremely weak correlation between the variables.
• Weight and runtime – positive insignificant association. Rho ≈ 0 with a 7% likelihood of association
• Weight and restpulse – negative insignificant association. Rho ≈ 0 with a 3% strength of association
• Weight and runpulse – positive insignificant association. Rho ≈ 0 with a 7% strength of association
• Weight and maxpulse – positive insignificant association. Rho ≈ 0 with a 14% strength of association
• Runtime and maxpulse – insignificant positive association of 21%

Weak association:

Variables with very weak associations (r or ρ between 0.25 and 0.5). Examples include age-runpulse and oxygen-maxpulse.

• Age and runpulse – Weak negative association. A decreased age might mean an increase in pulse rate while running. As the correlation is -0.29810, this means there is a 30% chance of this happening. And a 70% chance of the age having nothing to do with an individual’s pulse rate while running. Hence weak association.
• Age and maxpulse – weak negative association. Younger people (decrease in age) means more likelihood to have higher maximum pulse rates and vice versa. Rho is 0.38682 indicating a 39% chance of age and maximum pulse rate being associated. Hence weak association
• Oxygen and restpulse – weak negative association. Rho is -0.3828 which shows a weak relationship of 38% between the two variables.
• Oxygen and runpulse – weak negative association. 44% association between the two variables implying increased oxygen consumption results in decreased pulse rate while running.
• Oxygen and maxpulse – weak negative association of 32%
• Runtime and restpulse – weak positive association of 49%
• Runtime and runpulse – weak positive association of 28%
• Restpulse and runpulse – weak positive association of 37%.
• Restpulse and maxpulse – weak positive association of 33%

Strong association:

Only two variables show strong associations - runpulse and maxpulse (strong positive correlation) and oxygen and runtime (strong negative correlation). These findings confirm earlier observations.

Limitations of Chi-square test:

The chi-square test of association could not be performed due to the absence of categorical variables in the dataset. As such, the test is unreliable and not applicable to this dataset. In summary, as no categorical variables exist within the dataset, the chi square test of association is unreliable for this dataset. The chi square test is not exactly correlation analysis but for finding the relationship between categorical variables, which are nonexistent in this dataset.