Answers for the problem statement:

1.

a) Values in Column A is {25,35,21,67,98,27,64}

b) Values in Column B is {52,10,5,98,52,36,69}

c) Mean for Column A is given by the formula $\sum_{i=1}^{i=N} xi\ /N$ where N is the total No elements in the data set.

$\sum_{i=1}^{i=N} xi$= 25+35+21+67+98+27+64=337

$\bar{X}$=337/7=48.14

d) Mean for Column B is given by the formula $\sum_{i=1}^{i=N} yi\ /N$ where N is the total No elements in the data set.

e) $\sum_{i=1}^{i=N} yi$= 52+10+5+98+52+36+69=322

$\bar{Y}$=322/7=46

COV(x, y) = $1/(n-1)\sum_{i=1}^{i=N}(xi-\bar{x})(yi-\bar{y})$

As calculating covariance and correlation would be requiring a lot of intermediate values it would be simpler to represent all these values in a tabular form.

xi	yi	$(xi-\bar{x})$	$(yi-\bar{y})$	$(xi-\bar{x})(yi-\bar{y})$	$(xi-\bar{x})^2$	$(yi-\bar{y})^2$
25	52	-23.1429	6	-138.857	535.5918	36
35	10	-13.1429	-36	473.1429	172.7347	1296
21	5	-27.1429	-41	1112.857	736.7347	1681
67	98	18.85714	52	980.5714	355.5918	2704
98	52	49.85714	6	299.1429	2485.735	36
27	36	-21.1429	-10	211.4286	447.0204	100
64	69	15.85714	23	364.7143	251.449	529
				Total: 3303	Total:4984.85	Total:6382

f) Now n= 7

g) COV(x, y) = $1/(n-1)\sum_{i=1}^{i=N}(xi-\bar{x})(yi-\bar{y})$

h) 3303/6=550.5

i) The Correlation is defined as the ratio of COV(x, y) and the product of Std(x) and Std(y) where Std(x) is the standard deviation of x and Std(y) is the standard deviation of y.

j) Variance for the Column A: $S^2 = \sum_{i=1}^{i=n}(xi-\bar{x})^2/\ (n-1)$ = (4984.85/6) = 830.80

k) Standard Deviation for the Column A=$\sqrt[2]{S*S}$ = $\sqrt[2]{830.80}$ =28.82

l) Variance for the Column B: $S^2 = \sum_{i=1}^{i=n}(yi-\bar{y})^2/\ (n-1)$ = (6382/6) = 1063.66

m) Standard Deviation for the Column B=$\sqrt[2]{S*S}$ = $\sqrt[2]{1063.66}$=32.61

n) The Correlation coefficient = COV(x, y)/(std(x)*std(y)) = 550.5/(28.82*32.61)=

o) 550.5/(940.055)= 0.5857

2) The different way to deal with multi collinearity are as follows:

Multi collinear variable should be avoided before making the regression model as it inflate the model and it increase the variances between the coefficients which in turn makes the standard error between variable more significant .

Increased standard errors in turn means that coefficients for some independent variables may be found not to be significantly different from 0. In other words, by overinflating the standard errors, multi collinearity makes some variables statistically insignificant when they should be significant. Without multi collinearity (and thus, with lower standard errors), those coefficients might be significant.

a) Before making any regression model it is required to get rid of the redundant variable/highly collinear variable by using variable selection technique and avoiding the use of redundant variable by preparing the model.
b) We can ignore the collinear variable in case it is on the y axis that is dependent collinear variable does not affect the Regression model.

3) The correlation coefficient value or (ρ) value between two variable ranges from -1 to 1.

a) Where -1 indicate the very strong opposite relationship between the two variables that is if variable A increases then variable B decreases. And closer the value of ρ to -1 the more strong is the association between the two variables.
b) similarly if ρ=1 represent very strong direct relationship between the two variable A and B. That is if variable A increases then variable B also increases. And closer the value of ρ to 1 the more strong is the association between the two variables.
c) Whereas the value of ρ=0 indicate no association between the two variables.

Since there is always some degree of association/relationship between the variables by preparing the regression model we have to accept the some degree of correlation between the variables.

According to me the threshold value should lie in between -0.5>ρ<=-1 and 0.5<ρ>=1

4) The two type of variables used in ANOVA are as follow.

a) One Numerical Variable
b) One Categorical Variable

For Example we can use the ANOVA to compare the mean of two sample group such as you want to compare the average money holding (numerical variable) among three categories of the people(Rich, middle class and poor)(Categorical variable).

5) The Chi Square test is used to determine the independence of two nominal (categorical) variable of the same population.

Let us understand what is null and alternate Hypothesis in the Chi Square test with an example.

Suppose we have the two categorical variable in our population namely the gender of students participating in event which is also a categorical variable namely adventure and strategic.

And we have to identify whether the gender of students plays an important role in selection of the event category or not.

Since both the variable are the categorical variables we would be using the Chi Square test and will give propose the null and alternate hypothesis in such a way that both these are mutually exclusive and exhaustive.

So our Null hypothesis (H0) is: There is no relation between gender and the event selection

And our Alternate Hypothesis (Ha) is: There is a relationship between gender and the event selection.