DATA INTEGRATION EXERCISE

1) A food court manager wants to know if there is relationship between gender and the preferred condiment on burgers. The following table summerises the results. Test the hypothesis with significance level 10%

Condiment / Gender	Ketchup	Mustard	Relish	Total
Male	15	23	10	48
Female	2~~3~~5	19	8	52
Total	40	42	18	100

Ans. Significance level $\alpha = 0.1$

Hypothesis: H_0: Gender and Condiments are independent

H_a: Gender and Condiments are dependent

Degrees of freedom: $DF : (r-1)^*(c-1) = (2-1)^*(3-1) = 2$

Expected frequencies: $E_{r,c} = \frac{n_r{}^* n_c}{n_{Total}}$

Condiment / Gender	Ketchup	Mustard	Relish
Male	19.2	20.16	8.64
Female	20.8	21.84	9.36

Table for X^2:

Observed (O)	Expected (E)	$\frac{(O-E)^2}{E}$
15	19.2	0.91875
23	20.16	0.40008
10	8.64	0.21407
25	20.8	0.84808
19	21.8	0.36930
8	9.36	0.19761
	Total	2.94789

P-value for (2, 2.94789)
= 0.225

$\because p > 0.1$ the ~~reje~~ null hypothesis is accepted.

Gender and Condiments are independent.

2) Explain how data redundancy is handled in data integration.

Ans. During data analysis, various datastores are used for a given domain, which can lead to data redundancy. A data is said to be redundant if it can be derived from any other attribute or set of attribute. This can also be caused when there is an inconsistency in attribute or dimension naming. Handling redundancy involves identifying whether there is a dependency/dependencies among attributes. This is detected using the following methods:

- χ^2 Test: Used for nominal/categorical/qualitative data. The independence of the variables are tested.
- Correlation coefficent: Numerical/quantitative data is computed usually using Pearson's product moment coefficient. The higher the magnitude of the coefficient the stronger the correlation.

  $r = 0$ independent $r > 0$ directly proportional
  $r < 0$ indirectly proportional.

Once the dependencies are found, the dataset can be handled accordingly, i.e. unnecessary attributes may be removed.

3) Compare and contrast correlation and covariance

Correlation	Covariance.
When change in one results in change in other.	Mainly about direction in relationship between two variables. (positive or negative)
Strength of variables in comparison	Extent of change in a variable with respect to the other.
Corelation is scaled down covariance	Covariance is part of correlation
Value between 1 and -1	Any rational value.
Unit-free measure	Product of units of variables.
Zero correlation ensures independance.	Zero covariance doesn't necesarily mean independence