Concepts	Description
	Correlation and Regression
Covariance	**Covariance** : measure the linear relationship between 2 random variables $Cov_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X}) \times (Y_i - \bar{Y})}{n-1}$
Correlation coefficient	**Correlation coefficient** : strength of the linear relationship $r_{XY} = \frac{Cov_{X,Y}}{\sigma_X \times \sigma_Y}$ $-1 \leq r_{XY} \leq 1$
Scatter plot	**Scatter plot** : collection of point on the graph , each represents the value of 2 variables (X and Y) - Upward scatter plot : positive correlation - Downward scatter plot : negative correlation
Limitation to correlation analysis	**1. Outliers** : extreme values for sample observations → statistical evidence that significant relationship exist when there is none, or → no relationship when there is **2. Spurious Correlation** : may appear to have a relationship when there is none 3. Correlation only measure linear relationship, but not non-linear relationship
Test the correlation between 2 variables	Methodology : use t-test to test whether the correlation between 2 variables = 0 $H_o\colon \rho = 0\ ;\ H_a\colon \rho \neq 0$ $t = \frac{r \times \sqrt{n-2}}{\sqrt{1-r^2}}$ $df = n - 2$ $Reject\ H_0\ if\ t > +t_{critical}\ or\ t < -t_{critical}$
Dependent / Independent variables	**Dependent variables** : variable whose variation is explained by independent variables **Independent variables** : variable is used to explain the variation of dependent variables
Assumptions of linear regression	1. Linear regression exists between dependent and independent variables 2. Independent variable : uncorrelated with residuals 3. Expected value of residual term E(ε) = 0 4. Variance of the residual term is constant for all observations $E(\varepsilon_i^{\,2}) = \sigma_\varepsilon^{\,2}$ 5. Residual term is independently distributed (residual for observation A is not correlated with residual for observation B) 6. Residual term is normally distributed
Linear regression model	$Y_i = b_0 + b_1 \times X_i + \varepsilon_i$ In which : $Y = depedent\ variable$ $X = independent\ variable$ $b_1 = regression\ intercept\ term$ $b_0 = regression\ slope\ coefficient$ $\varepsilon = residual\ (error)$
Linear equation for regression line	$\widehat{Y_i} = \widehat{b_0} + \widehat{b_1} \times \widehat{X_i}$ In which : $\widehat{b_1} = estimated\ slope\ coefficient = \frac{Cov_{XY}}{\sigma_X^{\,2}}$ $\widehat{b_0} = estimated\ intercept\ term = \bar{Y} - \widehat{b_1} \times \bar{X}$ $\bar{Y} = mean\ of\ Y$ $\bar{X} = mean\ of\ X$
Confidence interval for regression slope coefficient (Range of b1)	Confidence interval of regression slope coefficient : $\widehat{b_1} \pm (t_c \times s_{\widehat{b_1}})$ or $\widehat{b_1} - (t_c \times s_{\widehat{b_1}}) < b_1 < \widehat{b_1} + (t_c \times s_{\widehat{b_1}})$ In which: $t_c = critical\ 2-tailed\ t\ value\ for\ the\ selected\ confidence\ level\ , with\ df = n-2$ $s_{\widehat{b_1}} = standard\ of\ error$
Test hypothesis that slope coefficient = hypothesized value	From the confidence interval for slope coefficient → use t-test to test the hypothesis that slope coefficient = hypothesised value $t_{b_1} = \frac{\widehat{b_1} - b_1}{s_{\widehat{b_1}}}$ $Reject\ H_0\ if\ t > +t_{critical}\ or\ t < -t_{critical}$