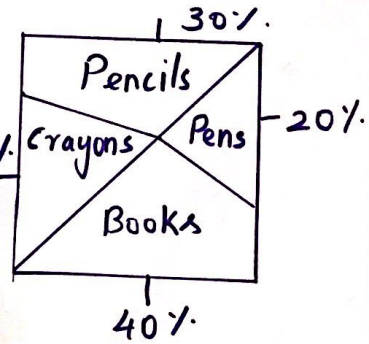
Data & Web Analytics — Math Problems COMP1810

1) A box that contains a total of 400 items were completely sold in an Online Shop and a chart was developed from the sale. Using the chart in percentages, calculate the number of each item that was originally in the box.

Solution:

Given Data



To find the no. of Items in each part of the box.

Pencils →

$\frac{30}{100} \times 400$

$= 120$ pencils

Pens → $\frac{20}{100} \times 400 = 80$

Crayons → $\frac{35}{100} \times 400$

$= 140$

Books → $\frac{40}{100} \times 400$

$= 160$

2) Using the below table, answer the following Questions

Age	Amount Spend Online (£)
23	100
44	87.3
27	97.8
	67.6
43	79.8
23	88.6
34	78.8
23	82.3
34	88.6
19	
43	7688.6
21	102.4
53	89.7

a) Identify the error in the Age Column. And Explain how to Correct the errors.

Ans:

There is one missing values in the Age Column. Hence all the other values don't have any outliers, the missing value can be replaced with mean of that Column.

So Skewness may be Zero or very least minimum Value

Replacing the mean value of Age Column

$$\frac{23+44+27+43+23+34+23+34+19+43+21+53}{\text{Total No. of Records}}$$

$$= \frac{387}{13} = 29.76 \approx \boxed{30}$$

b) Identify the errors in the "Amount Spend Online" column if any in the table and explain how to correct the errors with the justification.

Solution

The Error in the Column "Amount Spend Online" is 7688.6 which is Considered as the Outlier. The Outlier produced Skewness. So the value can be replaced with the median Value.

Median => the middle most Value of the given records.

Hence the Value as per given data is $\boxed{78.8}$

c) Identify the outliers in both Columns in the Given table if any found explain the Outlier.

Solution!

* There is no outlier in the "Age" Column

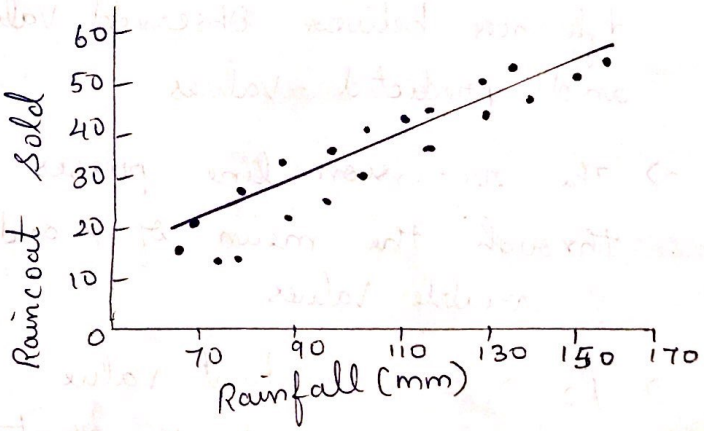
* The outlier found in the "Amount Spend Online" Column is 7688.6.

* An Outlier is any value or data that is Significantly different from the other set of values.

3) You are provided with the below graph representing the relationship between raincoat Sold and amount of rain fall. And it also be denoted by the equation,

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Explain each of the components in the above equation.



Solution:-

Y_i => Dependent Variable

β_0 => Y intercept

β_1 => Slope Coefficient

X_i => Independent Variable

ε_i => Random Residual Error Component.

Properties of Linear Regression Equation

-> The line in the given graph represents the "line of best fit" and it reduces the Sum of Squared differences between Observed Values and predicted Values.

-> The regression line passes through the mean of x and y variable values.

-> β_0 -> is a constant value which is equal to the y intercept of the graph.

-> The regression Co-efficient β_1 is the slope of the regression line which is equal to the average change in dependent Variable (y) for a unit change in the independent Variable (x).

-> Random Error (Residuals)

In regression the difference between the observed value of the dependent variable (Y_i) and the predicted value is called as Residuals.

$$\varepsilon_i = Y_{predicted} - Y_i$$

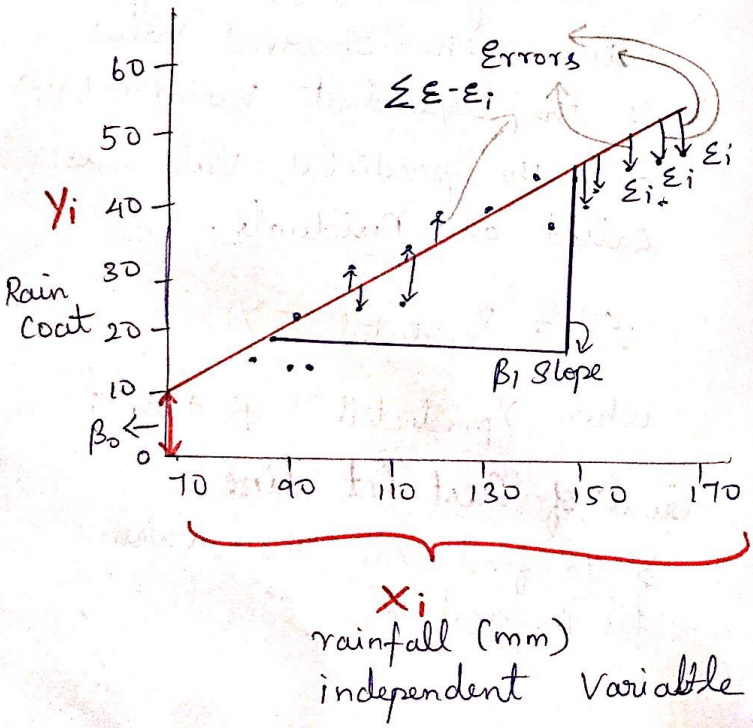
where $Y_{predicted} = \beta_0 + \beta_1 X_i$

Goal of Best fit Line.

-> To find the best values for β_0 and β_1.

→ The best fit line is a line that has the least error which means the error between predicted values and actual values should be minimum

values plotted based on Linear Regression



3.a) Illustrate and Justify that the above graph on the relationship between Rainfall and Raincoat Sold reveals,

"Correlation is not Causation
or
Correlation is Causation"

Solution:

→ The Graph obtained shows that the two variables Y and X are correlated but one does not cause the others.

→ Therefore the raise in the rainfall does not merely affect the Sales of the Rain Coat.

→ Here based on the plot there occurred perfect Positive Correlation, and Correlation is not a Causation.