Lecture 3 –EDA Fundamentals Continued

Q → Q

1) Does height affect salary?

   Dependent:
   Independent:

2) Simple Example – creating a scatterplot

   Plot the following:

Height (inches)	Income
72	$90,000
71	$86,000
68	$71,000
73	$89,000
70	$85,000
78	$100,000

There are 4 primary things to discuss pertaining to overall pattern when looking at a scatterplot:

1. Direction

2. Form

3. Strength (relative)

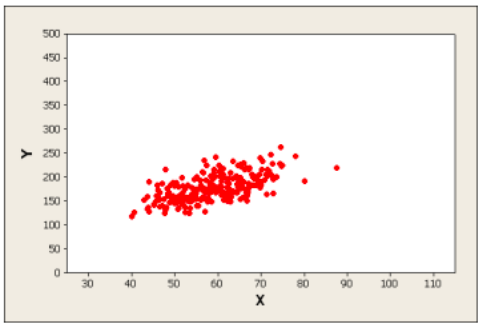
4. Deviations from the pattern (outliers)

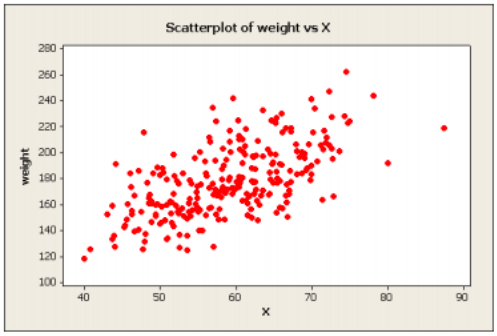
Magnitude Direction

Note: Use your intuition! Does the claimed relationship make sense?

Here we focus on linear relationships. So, how do you assess the strength of a linear relationship? First visually, but as always statisticians seek to quantify tendencies to validate.

For instance, we can always change the scale:





What do you notice?

So, how to we quantify such measures?

Correlation coefficient r –

a. Interpreting the value of r

i.

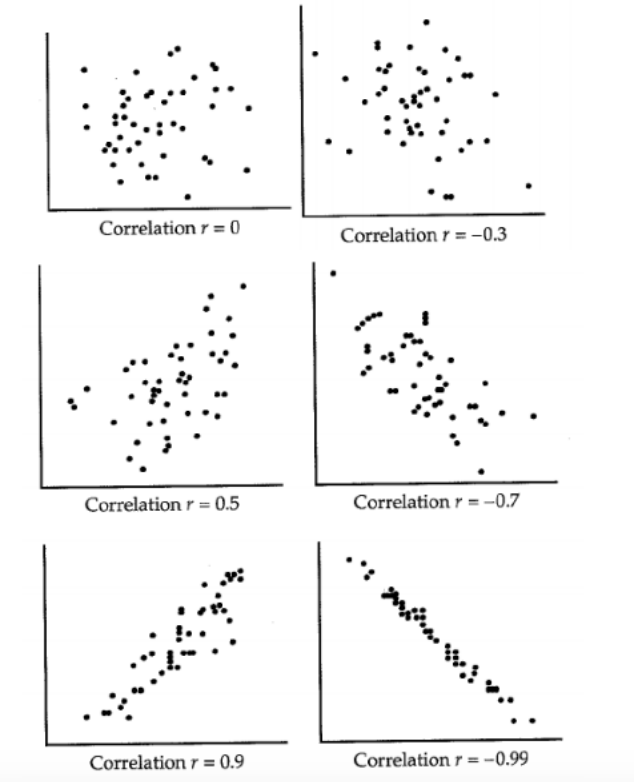
ii.

iii.

iv. r=0 implies that knowing one variable does not help in predicting the other

v. r=1 or r=-1 implies that knowing one variable will perfectly predict the other

vi. Examples – Getting a feel for strength



b. What are the cutoffs for “strong” and “weak”? Depends on the field of study and what you are used to seeing.

   i. In physical sciences r ≥ .90 is expected

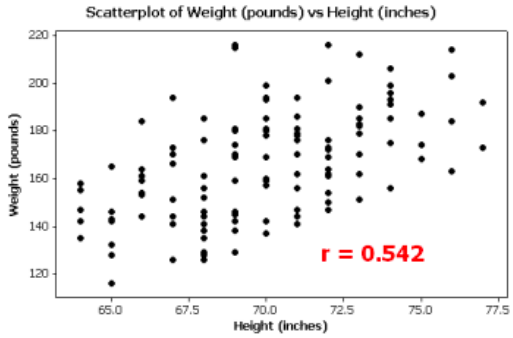
   ii. In psychology and social sciences, r of even 0.7 is considered excellent

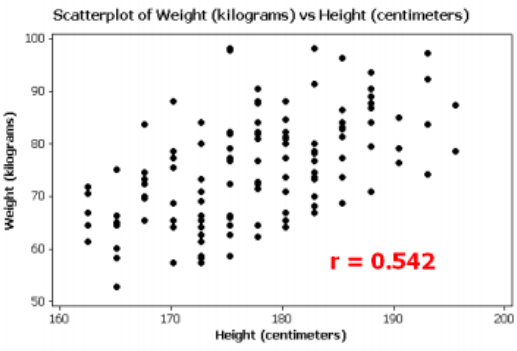
c. MUST CHECK THE SCATTERPLOT TO CONFIRM r MAKES SENSE

   1. Small r makes sense when it’s known that the explanatory is not the only variable that impacts the response.

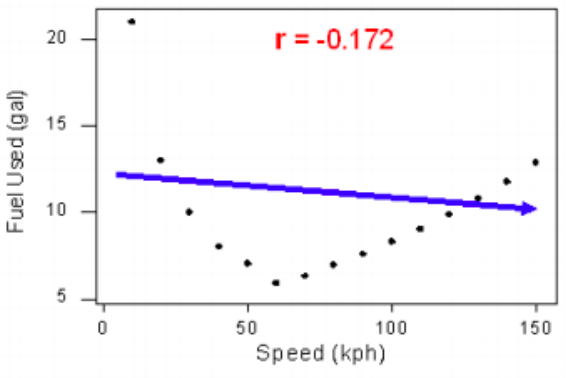
d. Properties of the correlation coefficient:

   1. If we change the unit we use for the response, explanatory, or both, this will not impact r (use intuition, would changing the unit impact the direction or strength of the relationship??)

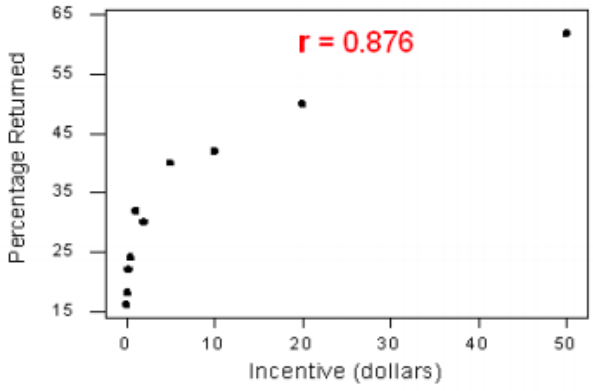




2. r is used for a **LINEAR** Relationship → you can get a value for r that is not 0 for a non-linear relationship. However, this value is useless.



Consequently, r is not enough by itself to claim the relationship is linear. We need the graph too.



3. r is sensitive to outliers of both magnitude and direction