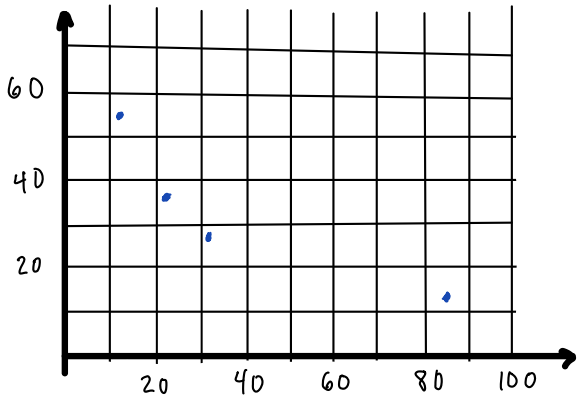
1. (a) There appears to be a negative linear association between variables $x$ and $y$ from this data set.




(b) We have

$$\bar{x} = \frac{1}{4}(12+85+31+22) = \frac{1}{4}(150) = 37.5$$

$$\bar{y} = \frac{1}{4}(54+13+28+37) = \frac{1}{4}(132) = 33$$

$$s_x^2 = \frac{1}{3}\left[(12-37.5)^2+(85-37.5)^2+(31-37.5)^2+(22-37.5)^2\right] = \frac{3189}{3} = 1063$$

$$\Rightarrow s_x \doteq 32.604$$

So, $b = (-0.9033)\frac{17.146}{32.604} \doteq -0.4750$, and

$$a = 33-(-0.475)(37.5) \doteq 50.813$$

giving us the line $\boxed{\hat{y} = -0.475x + 50.813}$.

(c) The desired residual is

$$\varepsilon = 28 - [(-0.475)(31)+50.81] = \boxed{-8.085}$$

2. (a) iii (b) ii (c) iv (d) i

3. (a) It is unimodal and right-skewed

(b) mean

(c) range $\approx 23-5 = 18$

5-number summary: 5, 8.5, 13, 19, 23

IQR $\approx 19-8.5 = 10.5$

4. resistant in this list are the mode, the median, and the IQR.

5. True statements: (ii), (iii) and (v)

6. (a) `tally( ~ dominantHand | sex, data = personalData)`

(b) `gf_histogram( ~ sqFt, data = houses)`

(c) `data(package = "mosaicData")`

(d) `sleep = c(8, 6.5, 7.5, 8.5, 8)`

(e) `names(houses)`

(f) `filter(houses, sqFt > 2000)`

7. We are given these probabilities, when a random message from the week is selected:

$P(\text{marked as spam}) = 0.101$

$P(\text{"free"}) = 0.041$

$P(\text{marked as spam and "free"}) = 0.0291$

(a) $$P(\text{"free"} \mid \text{marked as spam}) = \frac{P(\text{marked as spam and "free"})}{P(\text{marked as spam})}$$

$$= \frac{0.0291}{0.101} \doteq \boxed{0.288}$$

(b) $$P(\text{marked as spam} \mid \text{"free"}) = \frac{P(\text{marked as spam and "free"})}{P(\text{"free"})}$$

$$= \frac{0.0291}{0.041} \doteq \boxed{0.710}$$

| | Spam | Nonspam | |
|---|---|---|---|
| Free | 328 | 134 | 462 |
| No Free | 810 | 9995 | 10805 |
| | 1138 | 10129 | 11267 |

White

8. One can check independence of A and B by seeing if any one of these three equations hold:

i. $P(A \mid B) = P(A)$

ii. $P(B \mid A) = P(B)$

iii. $P(A \text{ and } B) = P(A)\,P(B)$

They all hold simultaneously, or none of them do.

As there are 1111 people in this group

$$P(A \text{ and } B) = \frac{55}{1111} \doteq 0.0495 \quad \text{and} \quad P(A)P(B) = \frac{103}{1111} \cdot \frac{740}{1111} \doteq 0.0618.$$

So, iii does <u>not</u> hold, and A, B are <u>not</u> independent.

9. (a) The study is observational in nature. No conditions are imposed on participants by the researchers.

(b) The cases are individuals from a single community, presumably all of them adolescents or teens.

(c) Is there an association between time spent watching TV and the number of aggressive acts committed?

(d) Explanatory variable (quantitative): time spent watching TV
Response variable (quantitative): number of aggressive acts committed

(e) Many possibilities here. For instance:

- level of parental supervision
- number of activities in which the individual participates as a teen