EXERCISE 1

a) Salary is a numerical variable, Department categorical. A suitable plot
is a multiple barplot with the conditional distribution of salary
for each department
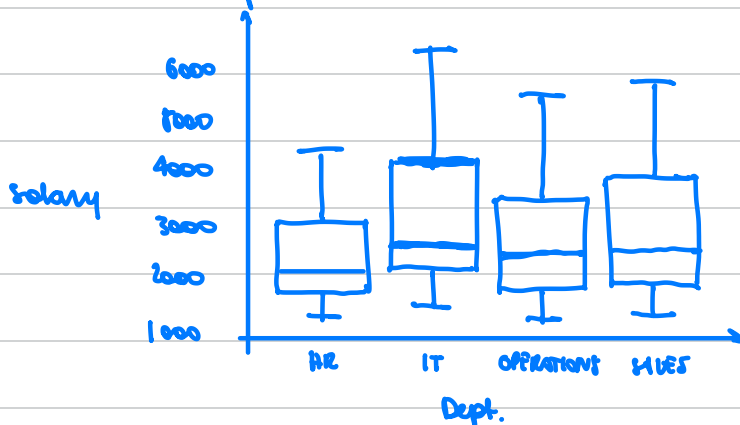
boxplot(Salary ~ Department)

There is a relationship between the two variables as the
distribution of salary changes depending on the department.
As an example, the median salary of HR employees is
below the first quartile of salary of IT employees.
Also IT employees can reach higher salaries.

b) Given a set of observations $x_1, x_2, \ldots, x_n$:

mean $:= \frac{1}{n} \sum_n x_i$

median is defined as the value exactly in the middle of
the ordered measurements if n is odd, the average
between the two central values if n is even.

One advantage of the mean is that it takes into account all values/measurements.
Also, the standard deviation is interpreted as the average distance from the mean.
One advantage of the median is that it is more robust w.r.t extreme values.

c) tapply (Salary, Department, median)

HR	IT	Operations	Sales
2117	2627	2467	2535

Median salary differ quite significantly between depts.
IT employees earn almost 25% more that HR employees.

d) quantile (Salary [Department == "IT"], c(0.85,1))
   between 4862 and 6404.

e) tab = table (Department, Role)
   prop.table (tab,1)

   the department with highest proportion of junior employees
   is IT, as the proportion of junior employees within each
   department is: HR: 45.98%   Operations: 52.83%
                    IT : 57.14%   Sales : 50.98%

EXERCISE 2

a) quantile (salary, c(0.25, 0.5, 0.75))

$Q_1 = 1972$ Median = 2504 $Q_3 = 3650$

b)

	[1000,1500]	[1500,2000]	(2000,3000]	(3000,5000]	(5000,8000]
Rel. Freq.	0.425	0.275	0.05	0.05	0.25
Freq. Densities	0.00085	0.00045	0.00005	0.000025	0.00008333

c) $Q_1 = 1000 + \dfrac{0.25}{0.00085} = 1294.118$

Median $= 1500 + \dfrac{0.5 - 0.425}{0.00045} = 1611.11$

$Q_3 = 5000$

note that if you collapse all observations in the middle point of the class they belong, the quartiles are completely off !!!

d) For Bill & Able:

mean = mean (salary) = 2868   sd = sd (salary) = 1180

For Conf/ult:

mean = sum (mids * p) = 2862

sd = sqrt (sum ((mids - mean)^2 * p)) = 2190

e) centrality while wrt means the centres of the two distributions are close, if we look at the median, salaries of confult are almost 100 lower!

dispersion salaries of Conf/ult are much more dispersed by looking at sd.

Shape by looking at mean, median and quartiles, both distributions are right skewed, but that of Conf/ult much more so!

Disparity... Bottom earners can be represented by those in the first quartile. Top earners by those in the last quartile

	First Quartile	Last Quartile
Bill & Able	< 1972	> 3650
Conf/ult	< 1294	> 5000

↳ Disparity is massively higher in conf/ult!

f) $UL = Q[3] + 1.5*(Q[3] - Q[1])$

Salary[Salary > UL]; Rate[Salary > UL]

The threshold is 6165. There are 2 outliers with salaries 6404 and 6360, both working in IT.

g) With respect to Conf/ult, looking at the table, the percentage earning above 6165 is:

$$\frac{(8000 - 6165)}{0.00008 3333} = 15.3\%$$ approximated assuming uniform distribution inside each class

Earning above 5000 is exactly those in the last class! So 25%, and it is an exact value.

EXERCISE 3

FALSE, TRUE, FALSE, TRUE, TRUE, TRUE

EXERCISE 4

a) For the 80% most typical days we should exclude those with 10% highest and 10% lowest employees working from home.

$$qpois(c(0.1, 0.9), 20)$$

Between 14 and 26 employees.

b) $1 - ppois(30, 20)$

$$P = 0.01347468$$

c) On average we expect this to happen

$$p \cdot 200 = 2.694836$$

If we model the number of times this happens as

$$Y \sim Poisson(p \cdot 200)$$

$$P(Y \leq 10) = ppois(10, p \cdot 200) = 0.9999 \quad (very\ high\ prob.!)$$

d) i) μ_2 should be set so that the average of 10 employees working from home is preserved.

$$0.2 \cdot 30 + 0.8 \cdot \mu_2 = 20 \implies \mu_2 = 17.5$$

ii) For 1 million days simulate if it rains or not

$$N = 1\ 000\ 000$$

$$rain = rbinom(N, 1, 0.2)$$

Then simulate the number of employees staying at home according to the suitable Poisson

$x = rain * rpois(N, 30) + (1-rain) * rpois(N, 17.5)$

now, in the most typical days we expect

quantile(x, c(0.1, 0.9))

between 13 and 30 employees.

iii) The probability of 30 or more employees working from home is:

mean($x > 30$) $\approx 0.108 = \hat{p}$

iv) finally

ppois($10, \hat{p} \cdot 200$) = 0.0046 (super low prob.!)