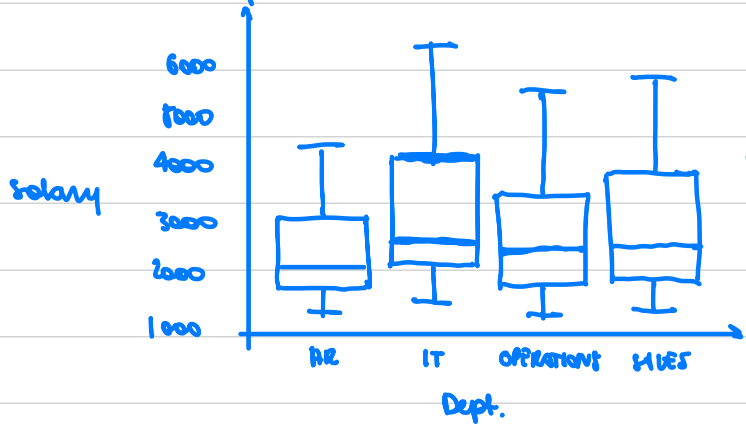
FIRST PARTIAL, VERSION A, FULL MARK PAPER

EXERCISE 1

a) Salary is a numerical variable, Department categorical. A suitable plot is a multiple boxplot 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 intended 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, mean)

HR	IT	Operations	Sales
2379	3121	2738	2981

Mean salary differ quite significantly between depts.
IT employees earn over 30% more that HR employees on average.

d) quantile(Salary[Department == "HR"], c(0.85, 1))
Between 3386 and 4440.

e) tab = table(Department, Role)
prop.table(tab, 1)
The department with highest proportion of senior employees is HR, as the proportion of senior employees within each department is:

HR	43.68%	Operations	33.96%
IT	27.27%	Sales	33.97%

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.	0.45	0.20	0.05	0.05	0.25
Freq. Densities	0.0009	0.0004	0.00005	0.000025	0.00008333

c) $Q_1 = 1000 + \frac{0.25}{0.0009} = 1277.78$

Median $= 1500 + \frac{0.5 - 0.45}{0.0004} = 1625$

$Q_3 = 5000$

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 Con#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 conHult are almost 100 lower!

dispersion salaries of Con#ult are much more dispersed by looking at sd.

Shape by looking at mean, median and quantiles, both distributions are right skewed, but that of Con$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
Con$ult	< 1277	> 5000

↳ Disparity is massively higher in con$ult!

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

Salary[Salary > UL]; Role[Salary > UL]

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

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

$$\frac{(8000-6165)}{0.000083333} = 15.3\%$$ 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, FALSE, TRUE, FALSE, FALSE, FALSE

EXERCISE 4

a) For the 90% most typical days we should exclude those with 5% highest and 5% lowest employees working from home.

qpois(c(0.05,0.95), 20)

Between 13 and 28 employees.

b) 1 - ppois(29, 20)

$p = 0.0218182$

c) On average we expect this to happen $p \cdot 200 = 4.363644$ times.

If we model the number of times this happens as $Y \sim \text{Poisson}(p \cdot 200)$

$P(Y \leq 10) =$ ppois(10, p·200) $= 0.9946$ (very high prob.!)

d) i) μ_2 should be set so that the average of 20 employees working from home is preserved.

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

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

N = 1000 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(0.05, 0.95))

between 11 and 34 employees.

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

mean(x >= 30) ≈ 0.108 = $\hat{p}$

iv) Finally

ppois(10, $\hat{p}$·200) = 0.004385 (super low prob.!)