Wesley Roots

# MAST7866 – Foundations of Data Science

# Assessment 1

Working out shown in non-bold black text

Final answer shown in non-bold dark blue text

## 1.

### (a)

$P(A \cup B \cup C)$

### (b)

$(A \cap B' \cap C') \cup (A' \cap B \cap C') \cup (A' \cap B' \cap C)$

### (c)

$(A' \cap B \cap C) \cup (A \cap B' \cap C) \cup (A \cap B \cap C')$

### (d)

$(A \cap \mathrm{B} \cap \mathrm{C})'$

## 2.

### (a)

- (i) $\frac{5}{11}$
- (ii) $\frac{9}{11}$
- (iii) $\frac{7}{11}$

### (b)

`1 - (pnorm(45.04, 45, 0.04) - pnorm(44.94, 45, 0.04))` → `0.2254625`

Therefore, probability = 22.55%

## 3.

### (a)

Given $A \cap C = 0, B \cap \mathrm{C} = 0, P(A \cup B) = 1$

$\mathbf{1} = \boldsymbol{P}(\boldsymbol{A} \cup \boldsymbol{B} \cup \boldsymbol{C})$

$= \boldsymbol{P}(\boldsymbol{A}) + \boldsymbol{P}(\boldsymbol{B}) + \boldsymbol{P}(\boldsymbol{C}) - \boldsymbol{P}(\boldsymbol{A} \cap \mathrm{B}) - \mathrm{P}(\mathrm{A} \cap \mathrm{C}) - \mathrm{P}(\mathrm{B} \cap \mathrm{C}) + \mathrm{P}(\mathrm{A} \cap \mathrm{B} \cap \mathrm{C})$

Therefore:

$1 = \frac{1}{3} + \frac{1}{4} + \frac{1}{2} - P(A \cap \mathrm{B})$

$P(A \cap \mathrm{B}) = \frac{1}{12}$

As $P(A) \cdot P(B) = \frac{1}{12}, P(A) \cdot P(B) = P(A \cap \mathrm{B})$

### (b)

$P(A \cup C) = P(A) + P(C) - P(A \cap C)$
As $P(A \cap C) = 0: P(A) + P(C)$

## 4.

`(d) sort`

## 5.

### (a)

```
suburban <- c(58.5, 60.8, 60.6, 64.3, 64.1, 40.7, 43.7, 48.6, 49.3,
49.5)
median(suburban)
```

Median = 54

### (b)

```
urban <- c(97.8, 68.3, 109.2, 78.1, 113.7, 78.6, 122.0, 84.4, 125.5,
85.3)
round(IQR(urban))
```

IQR = 33

**(c)**

```
sd(suburban)
sd(urban)
```

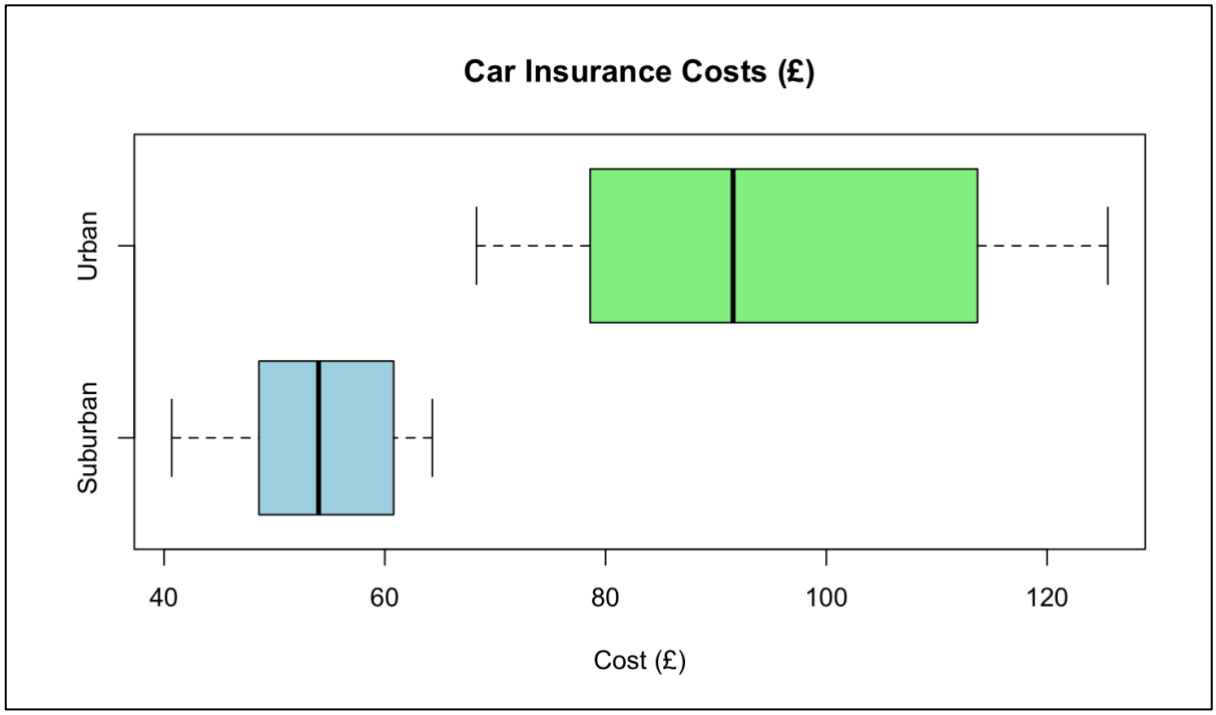
Suburban STDev = 8.65
Urban STDev = 20.21

Urban insurance costs vary much more than suburban costs due to a significantly larger standard deviation

**(d)**

```
boxplot(suburban, urban,
      names = c("Suburban", "Urban"),
      main = "Car Insurance Costs (£)",
      xlab = "Cost (£)",
      col = c("lightblue", "lightgreen"),
      horizontal=TRUE)
```

Output:




Urban costs vary significantly more than suburban costs, exhibited by the larger IQR. Urban costs are also slightly right-skewed while suburban costs are roughly symmetric with a very slight left-skew. Using the 1.5x IQR rule, there are no outliers in the suburban or urban group.

**(e)**

- The car insurance prices are higher in urban areas than suburban areas: True
- There are outliers in the urban dataset: False

## 6.

**(a)**

```
forbes <- read.table("forbes.txt", header = TRUE)
attach(forbes)
```

**(b)**

- (i)
  ```
  length(age) [1] 40
  length(salary) [1] 40
  length(degree) [1] 40
  ```
- (ii)
  ```
  sum(age<55) [1] 8
  sum(salary[age<55])/sum(age<55) [1] 33.94
  sum(salary|age<55)/sum(age<55) [1] 5
  ```
- (iii)
  ```
  salary/age
  [1] 4.5349091 3.6595238 2.1532308 2.2588333 2.0445000
  1.1370423 1.2349180 1.2603509 1.2221053 1.0292537 1.1248214
  [12] 0.9736207 0.9331667 0.9964583 0.7450000 0.6355385
  0.6292063 0.7418868 0.6814545 0.7060377 0.6193220 0.6063333
  [23] 0.6001695 0.5946552 0.7660465 0.5114062 0.5647368
  0.5006349 0.4912500 0.4988889 0.5311864 0.5119672 0.5415789
  [34] 0.5539623 0.5337037 0.3944444 0.5179630 0.5578000
  0.4483333 0.4061290
  ```

**(c)**

```
min(age) [1] 43
max(age) [1] 72
mean(age) [1] 59.03
min(salary) [1] 25.18
max(salary) [1] 249.42
mean(salary) [1] 58.31
```

**(d)**

```
deg <- table(tools::toTitleCase(degree))
```

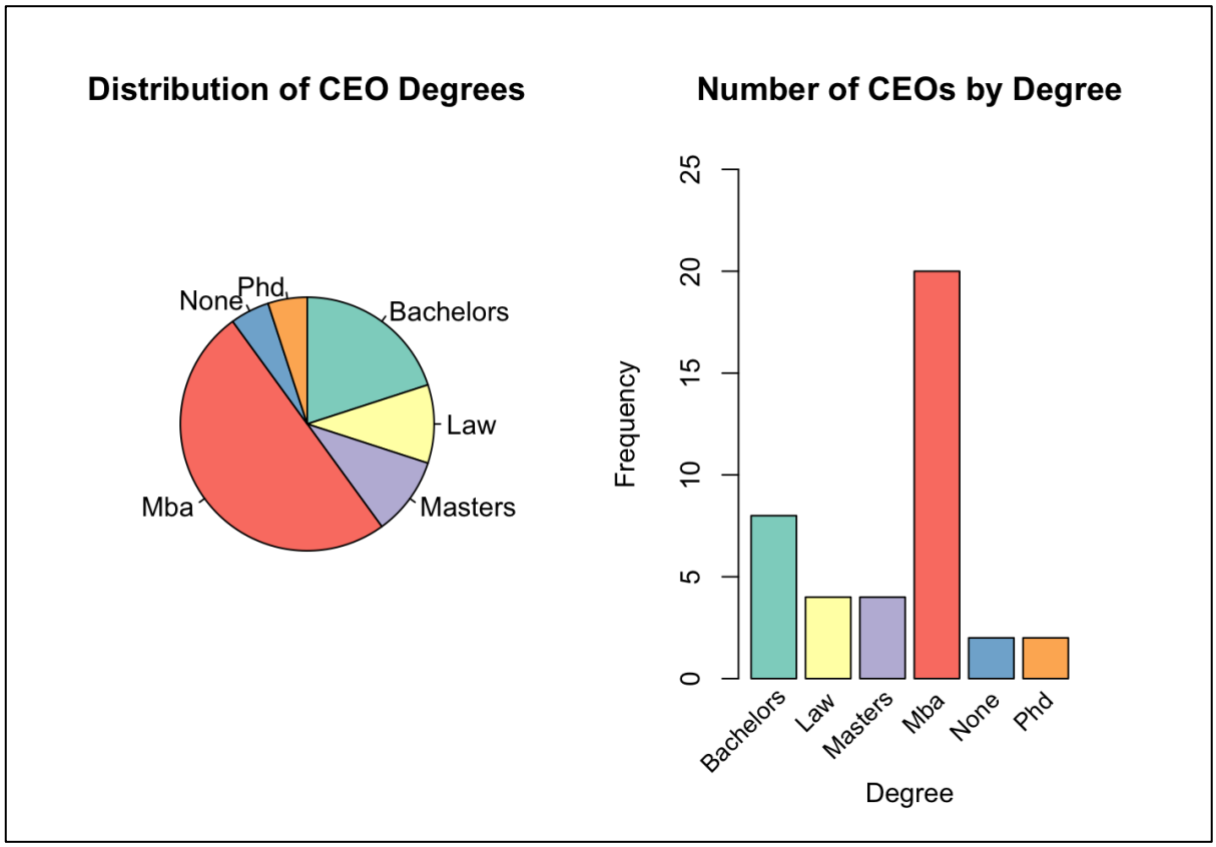
**(e)**

```
par(mfrow = c(1,2), mar = c(4,4,4,4))
pie(deg,
    main = "Distribution of CEO Degrees",
    col = RColorBrewer::brewer.pal(length(deg), "Set3"),
    clockwise = TRUE,
)

axis1 <- barplot(deg,
        main = "Number of CEOs by Degree",
        col = RColorBrewer::brewer.pal(length(deg), "Set3"),
        ylim = c(0,25),
        xlab = "Degree",
        ylab = "Frequency",
        xaxt ="n"
)

text(x = axis1,
     y = -1,
     labels = names(deg),
     srt = 45,
     adj = 1,
     xpd = TRUE,
     cex = 0.9)
```

Output:



From the above graph, most CEOs have advanced degrees such as an MBA, master's degree, law degree, or PhD. Only a small percentage hold a Bachelors (~8%) or no degree (~2.5%).

**(f)**

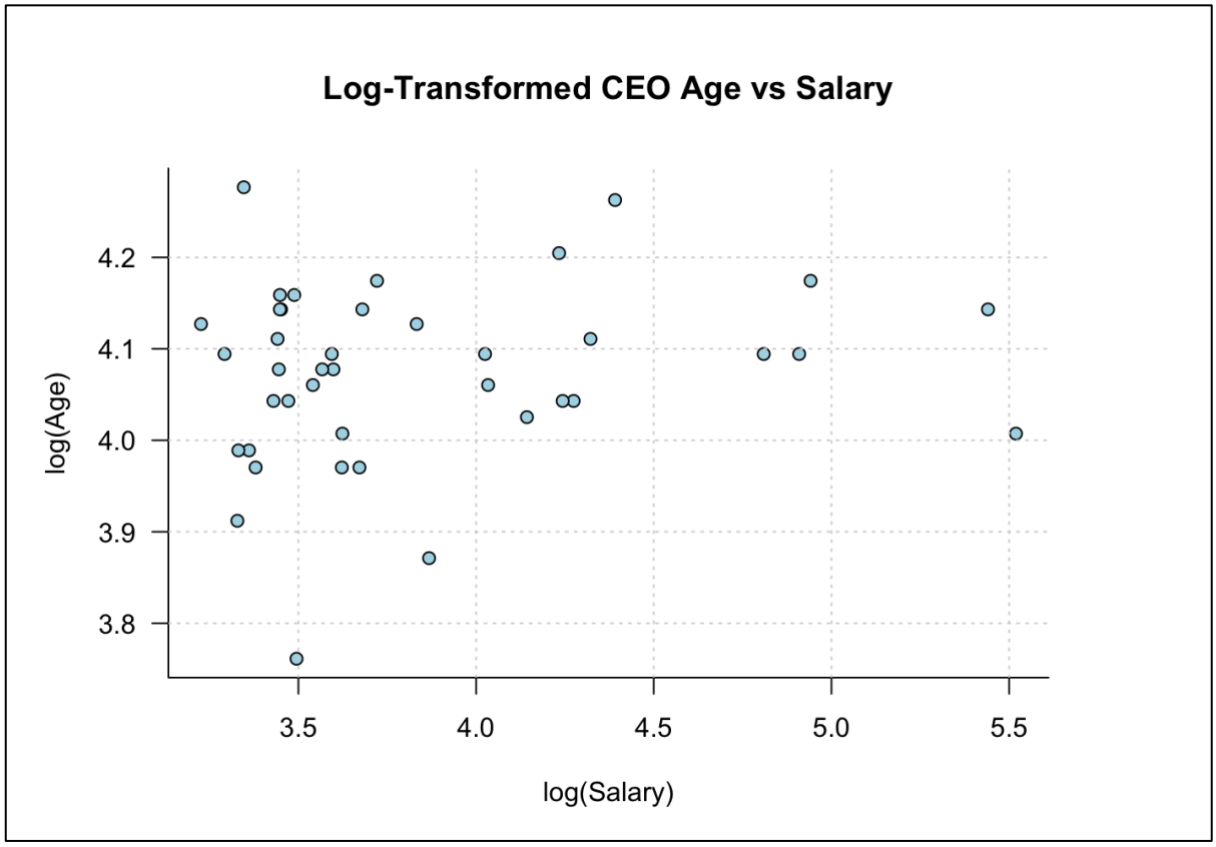
```
par(mfrow=c(1,1), mar = c(5,5,5,5))
plot(forbes$salary, forbes$age,
     main = "CEO Age vs Salary",
     xlab = "Salary",
     ylab = "Age",
     pch = 21,
     bg = "grey",
     ylim = c(30,80),
     las = TRUE,
     bty = "l"
)
grid()
```

Output:


CEO salaries and ages shows a weak correlation. Most salaries range between 25-75 for CEOs between the ages of 50 and 65, with a few higher-salary outliers.

The distribution shows a non-linear pattern.

**(g)**




The log-transformed plot shows a weak linear relationship. Most CEO ages fall between 3.95 and 4.2, while log(salary) is more spread out, with a few higher-salary outliers.

## 7.

```
bill <- c(49,36,42,43,49,30,48,32,49,35,30,48)

sum(bill)
[1] 491
min(bill)
[1] 30
max(bill)
[1] 49

over_41 <- sum(bill > 41)
over_41 / length(bill) * 100
[1] 58.33333
```

Therefore, 58.33% of monthly bills were higher than 41