Project 6

PROBABILITY ANALYSIS AND EDA

Aditya Parey

ALY 6000 Summer 2024

Overview and Key Findings:

The outcomes of several probabilistic and exploratory data analysis tasks performed on various datasets and scenarios are presented in this study. The examination consists of:

Probability of Winning Games: Using simulations, determine the likelihood of winning precisely ten games.

Failure Analysis of Light Bulbs: Using a Poisson distribution, model the number of failures in a given time frame.

Analysing the quantity of clients who arrive at a call centre is known as call centre arrival analysis.

Exploration of the Penguins Dataset: This involves using the palmerpenguins package's penguins dataset for exploratory data research.

The report outlines the methodologies employed, the principal discoveries, and the conclusions drawn from the investigation.

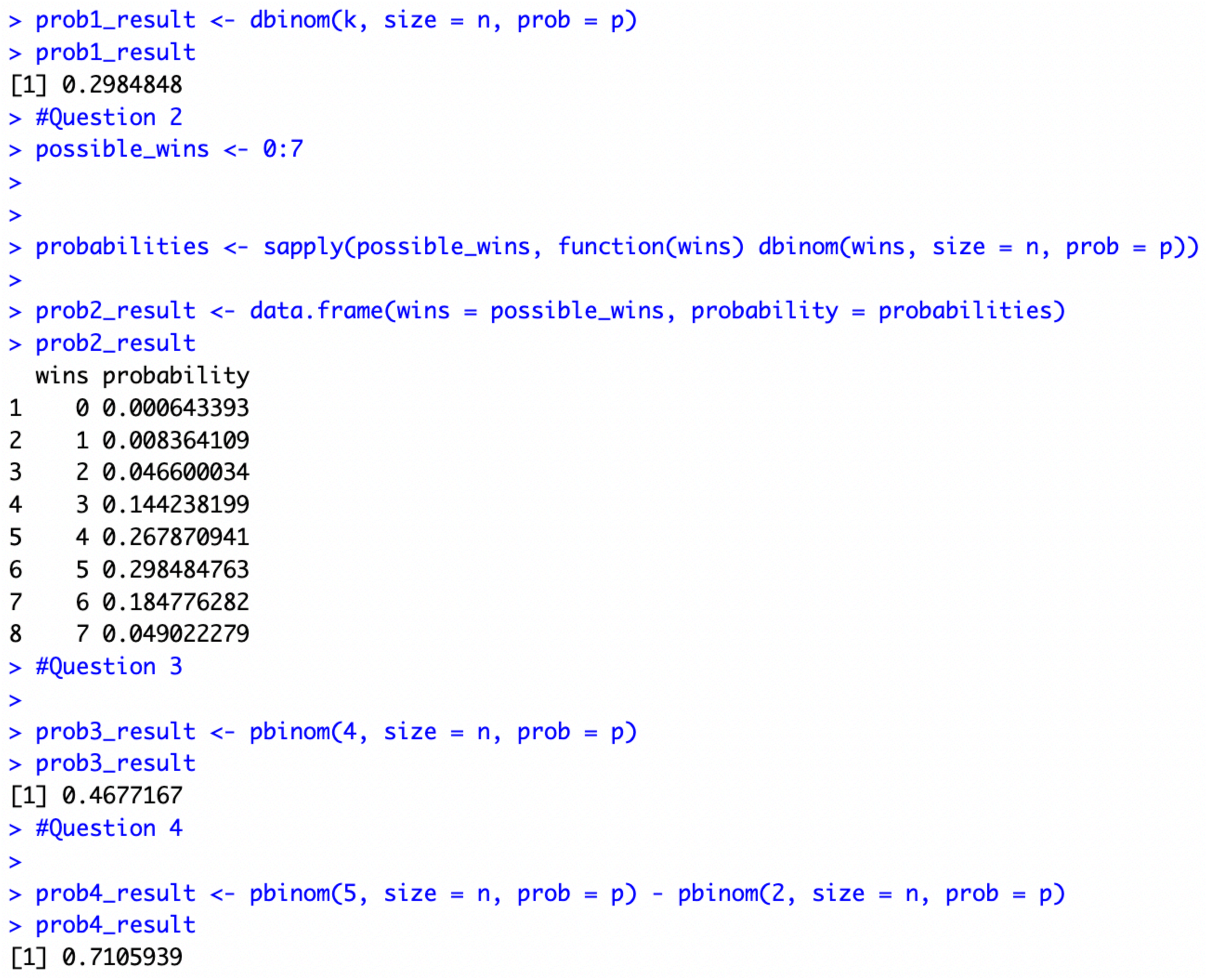
Probability of winning games:

Goal: Using simulation methods, determine the likelihood that a soccer club will win precisely ten games in a series.

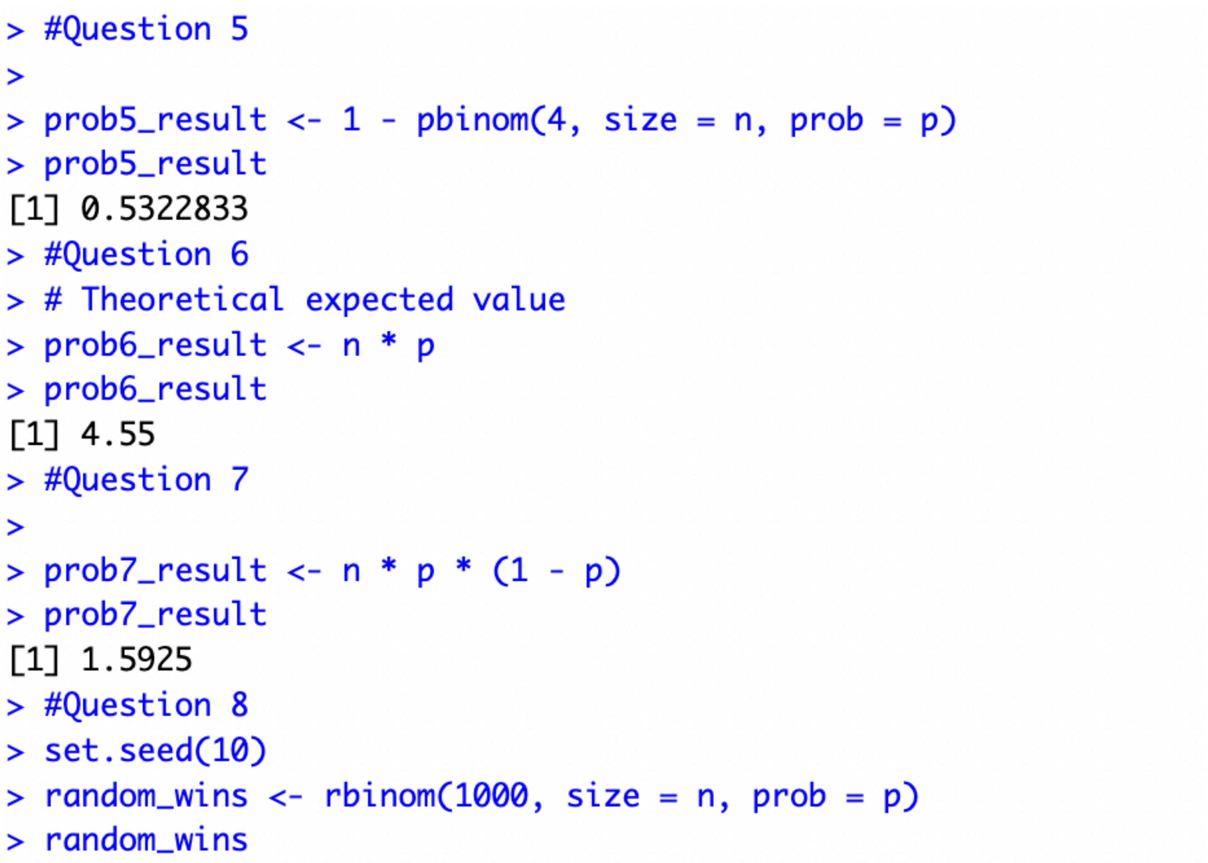
Techniques:

To simulate the likelihood of winning precisely ten games, a bespoke simulation function called mysim() and a loop-based methodology were employed.
The probability was estimated by running several simulations using the replicate() method.
Principal Results:

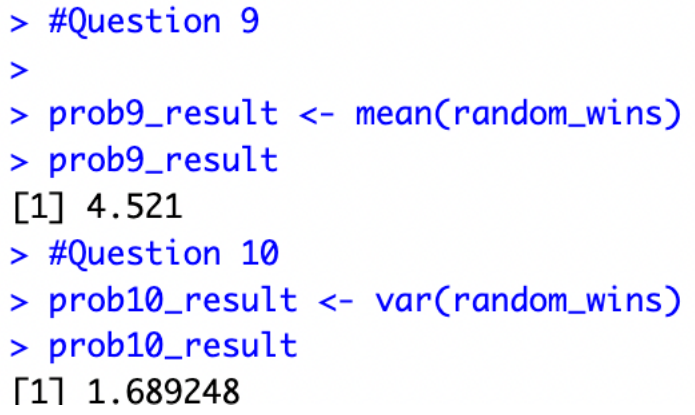
According to the simulation, the team's consistency in performance as well as the probability of finishing the season with a particular number of victories.



```
> prob1_result <- dbinom(k, size = n, prob = p)
> prob1_result
[1] 0.2984848
> #Question 2
> possible_wins <- 0:7
>
>
> probabilities <- sapply(possible_wins, function(wins) dbinom(wins, size = n, prob = p))
>
> prob2_result <- data.frame(wins = possible_wins, probability = probabilities)
> prob2_result
  wins probability
1    0 0.000643393
2    1 0.008364109
3    2 0.046600034
4    3 0.144238199
5    4 0.267870941
6    5 0.298484763
7    6 0.184776282
8    7 0.049022279
> #Question 3
>
> prob3_result <- pbinom(4, size = n, prob = p)
> prob3_result
[1] 0.4677167
> #Question 4
>
> prob4_result <- pbinom(5, size = n, prob = p) - pbinom(2, size = n, prob = p)
> prob4_result
[1] 0.7105939
```



```
> #Question 5
>
> prob5_result <- 1 - pbinom(4, size = n, prob = p)
> prob5_result
[1] 0.5322833
> #Question 6
> # Theoretical expected value
> prob6_result <- n * p
> prob6_result
[1] 4.55
> #Question 7
>
> prob7_result <- n * p * (1 - p)
> prob7_result
[1] 1.5925
> #Question 8
> set.seed(10)
> random_wins <- rbinom(1000, size = n, prob = p)
> random_wins
```



```
> #Question 9
>
> prob9_result <- mean(random_wins)
> prob9_result
[1] 4.521
> #Question 10
> prob10_result <- var(random_wins)
> prob10_result
[1] 1.689248
```

Analysis of Light Bulbs Failure

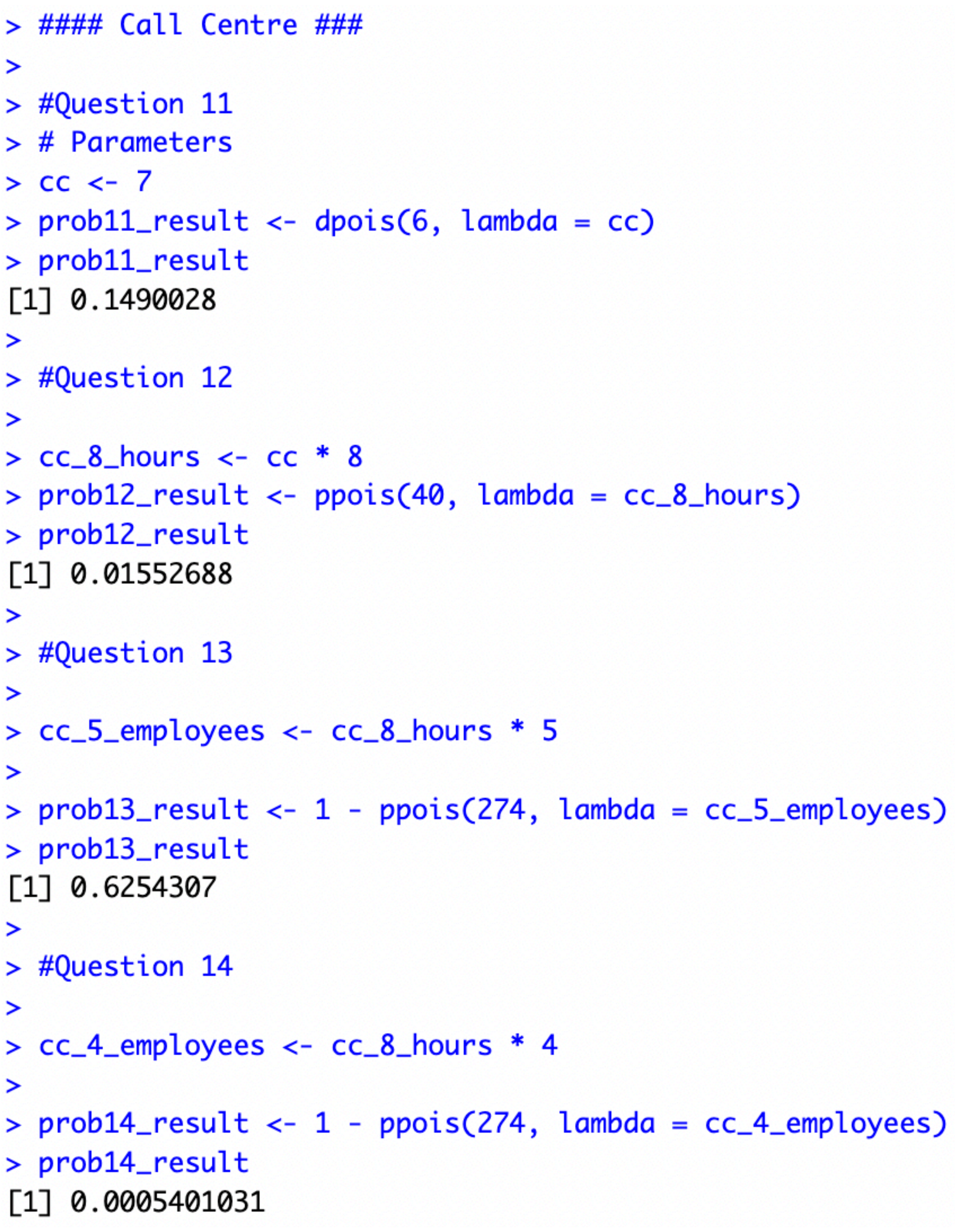
The goal is to use the Poisson distribution to model the number of light bulb failures that occur over a given period of time.

Techniques:

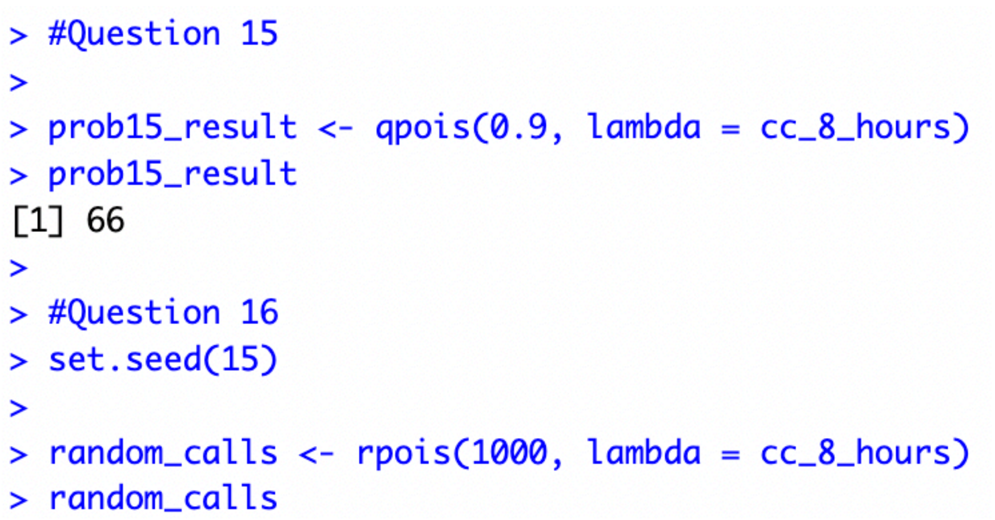
The Poisson distribution was used to calculate the likelihood of a specific number of failures given a mean (cc).

Principal Results:

Situations like light bulb failure, where events happen separately over a set period of time, are well-modeled by the Poisson distribution.



```
> #### Call Centre ###
>
> #Question 11
> # Parameters
> cc <- 7
> prob11_result <- dpois(6, lambda = cc)
> prob11_result
[1] 0.1490028
>
> #Question 12
>
> cc_8_hours <- cc * 8
> prob12_result <- ppois(40, lambda = cc_8_hours)
> prob12_result
[1] 0.01552688
>
> #Question 13
>
> cc_5_employees <- cc_8_hours * 5
>
> prob13_result <- 1 - ppois(274, lambda = cc_5_employees)
> prob13_result
[1] 0.6254307
>
> #Question 14
>
> cc_4_employees <- cc_8_hours * 4
>
> prob14_result <- 1 - ppois(274, lambda = cc_4_employees)
> prob14_result
[1] 0.0005401031
```



```
> #Question 15
>
> prob15_result <- qpois(0.9, lambda = cc_8_hours)
> prob15_result
[1] 66
>
> #Question 16
> set.seed(15)
>
> random_calls <- rpois(1000, lambda = cc_8_hours)
> random_calls
```

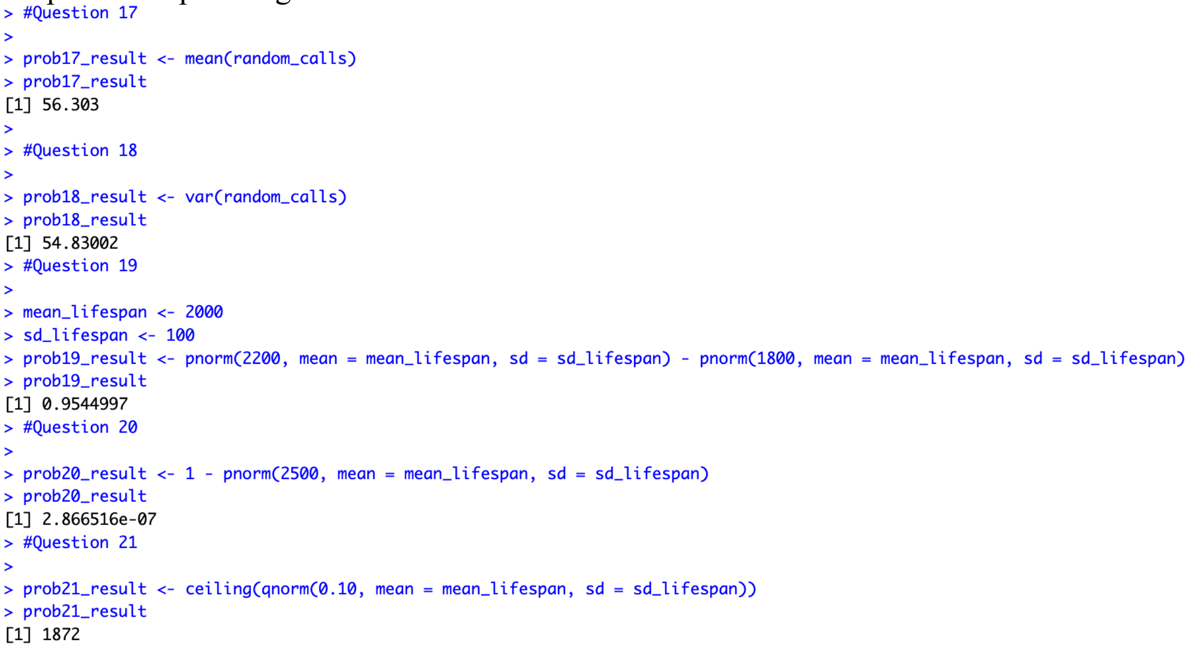
Arrival Analysis of Call Centres:
(Question 19 – 27)

The goal is to use a probabilistic model to analyse the number of customers who arrive at a call centre.

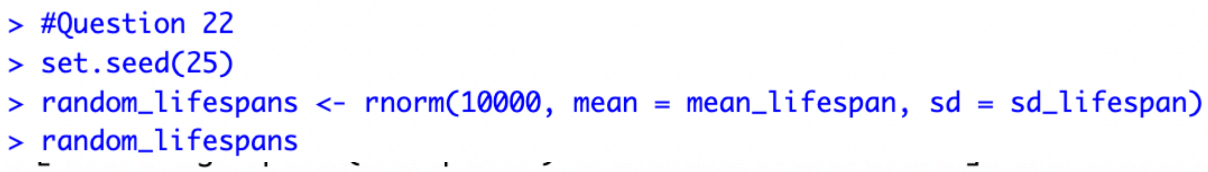
Techniques:

The Poisson distribution was used to tackle the challenge of estimating the number of arrivals in a specified amount of time.
The likelihood of various numbers of arrivals was calculated using the expected number of arrivals.

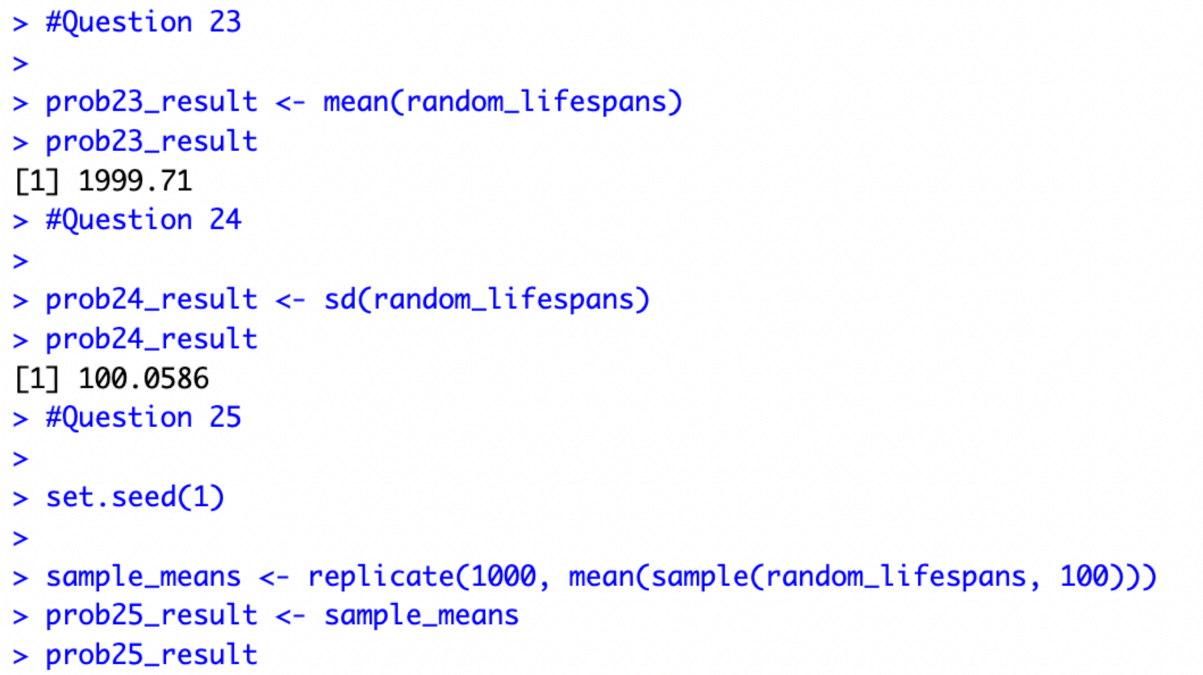
Principal Results:
Call centre operations can be optimised to cut down on wait times by using this information for personnel planning.



```
> #Question 17
>
> prob17_result <- mean(random_calls)
> prob17_result
[1] 56.303
>
> #Question 18
>
> prob18_result <- var(random_calls)
> prob18_result
[1] 54.83002
> #Question 19
>
> mean_lifespan <- 2000
> sd_lifespan <- 100
> prob19_result <- pnorm(2200, mean = mean_lifespan, sd = sd_lifespan) - pnorm(1800, mean = mean_lifespan, sd = sd_lifespan)
> prob19_result
[1] 0.9544997
> #Question 20
>
> prob20_result <- 1 - pnorm(2500, mean = mean_lifespan, sd = sd_lifespan)
> prob20_result
[1] 2.866516e-07
> #Question 21
>
> prob21_result <- ceiling(qnorm(0.10, mean = mean_lifespan, sd = sd_lifespan))
> prob21_result
[1] 1872
```

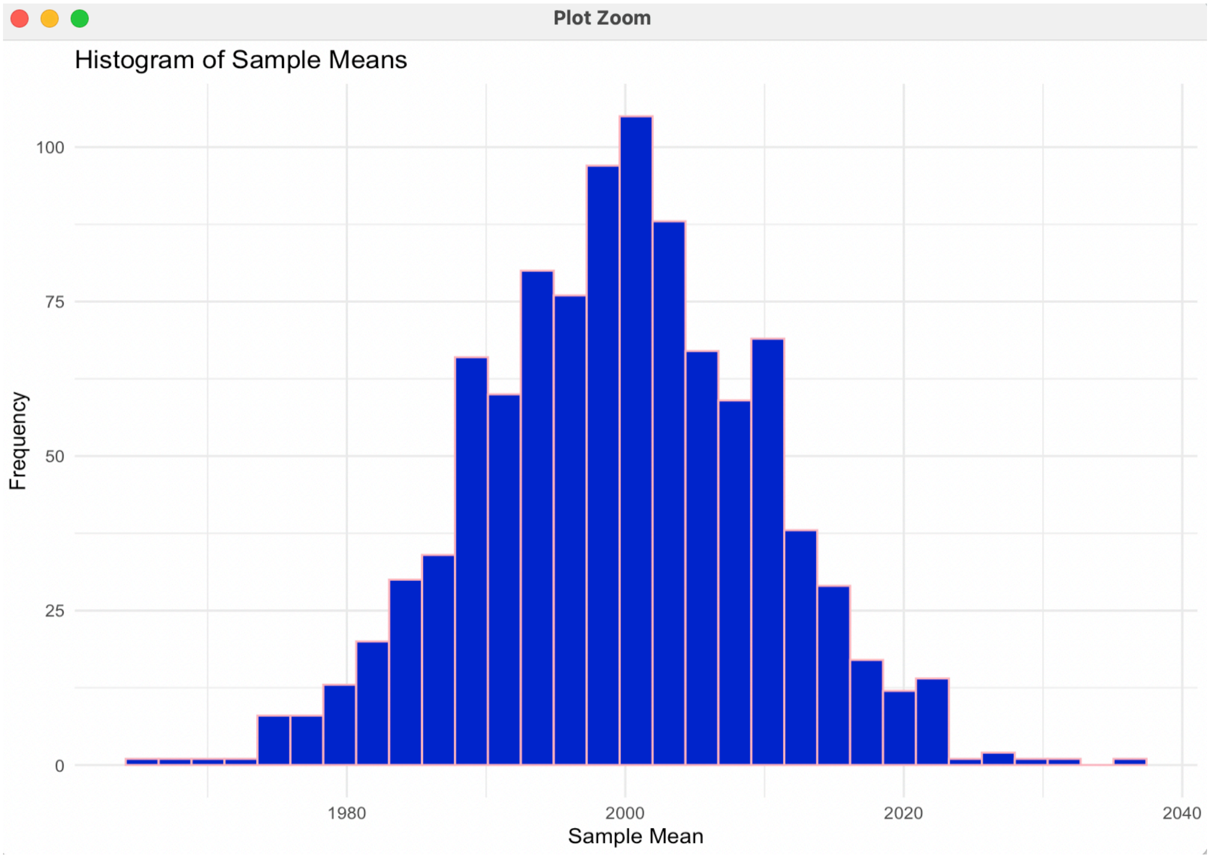


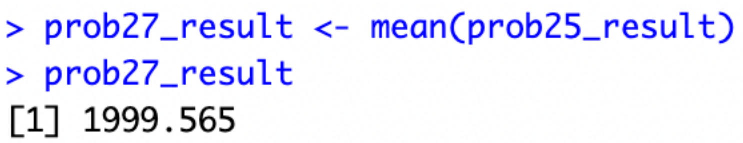
```
> #Question 22
> set.seed(25)
> random_lifespans <- rnorm(10000, mean = mean_lifespan, sd = sd_lifespan)
> random_lifespans
```



```
> #Question 23
>
> prob23_result <- mean(random_lifespans)
> prob23_result
[1] 1999.71
> #Question 24
>
> prob24_result <- sd(random_lifespans)
> prob24_result
[1] 100.0586
> #Question 25
>
> set.seed(1)
>
> sample_means <- replicate(1000, mean(sample(random_lifespans, 100)))
> prob25_result <- sample_means
> prob25_result
```

Question 26)





```
> prob27_result <- mean(prob25_result)
> prob27_result
[1] 1999.565
```

Penguins Dataset Investigation

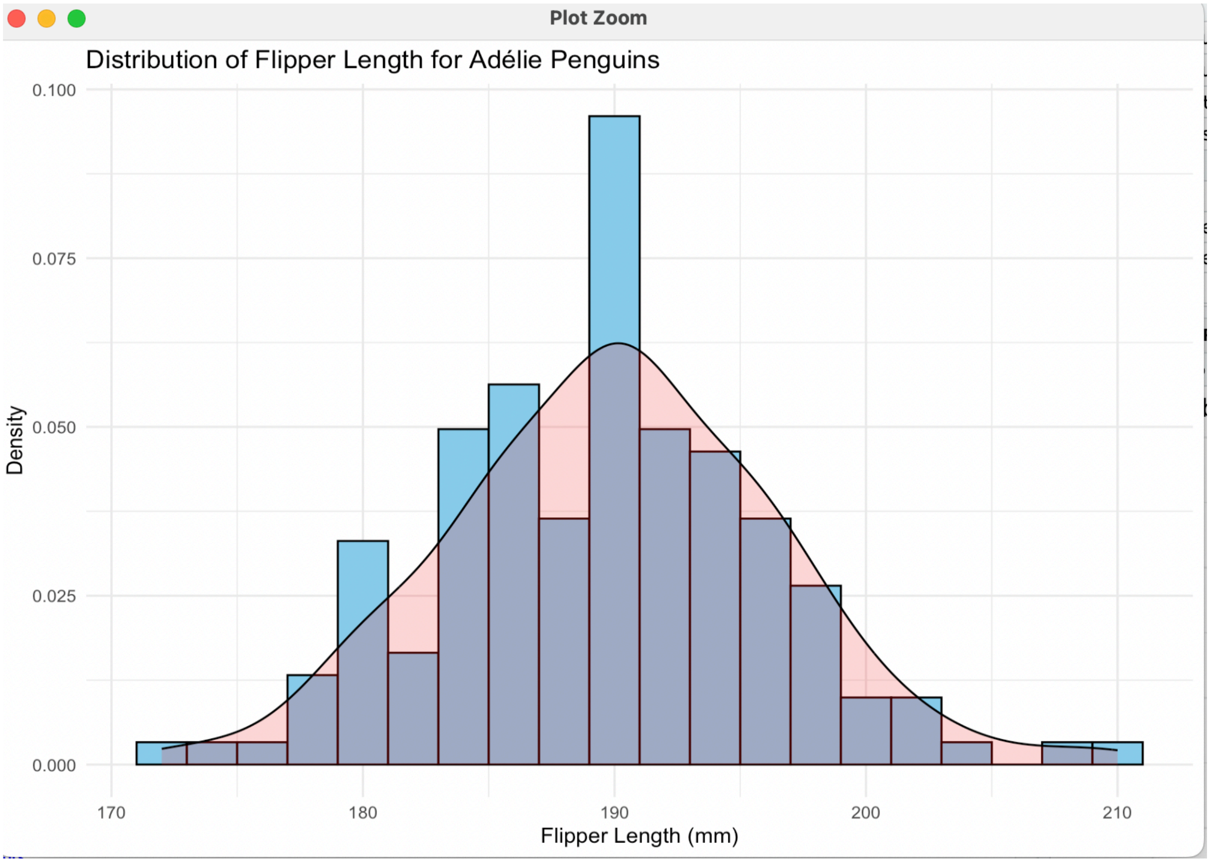
The goal is to conduct an exploratory data analysis (EDA) using the palmerpenguins package's penguins dataset.

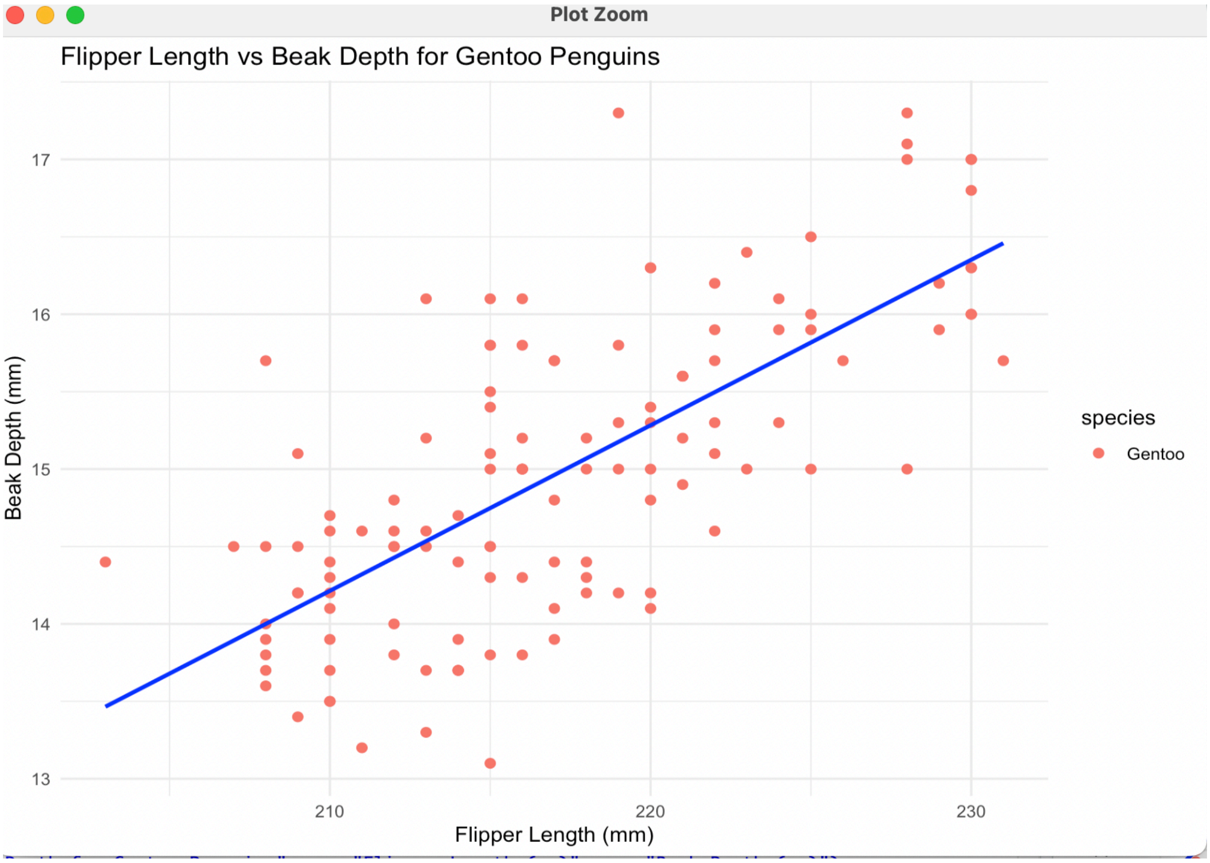
Techniques:

To investigate the correlations between variables including species, bill length, bill depth, and flipper length, summary statistics, visualisations, and cross-tabulations were produced.
The data were shown using a variety of charts, such as boxplots and scatterplots.

Principal Results:

Variations by Species: The length, depth, and flipper length of each species were found to differ significantly.
Positive associations between bill length and flipper length were found, indicating that longer bills are associated with longer flippers in penguins.
Visualisations: The distribution and variability among various species were clearly shown by the boxplot of flipper length by species.





Analysis of Probability Test Cases:

Game Strategies: The likelihood of winning precisely ten games provides information on the soccer team's strategic planning. When establishing objectives and goals for the season, the team management should take this probability into account.

Planning for Electricity Maintenance: To avoid disruptions, the failure analysis of light bulbs recommends putting in place a maintenance program based on the anticipated number of failures.

Staff Optimisation: Based on the arrival analysis of call centres, it is evident that dynamic staffing models are necessary to improve customer service efficiency by adapting to the anticipated number of arrivals.
Penguins Information Set:

Conservation initiatives: By highlighting the distinctive physical traits of each species, the insights from the penguins dataset can support conservation initiatives that are species-specific.

References:

- Penguins Dataset: Horst AM, Hill AP, Gorman KB. palmerpenguins: Palmer Archipelago (Antarctica) penguin data. R package version 0.1.0. https://allisonhorst.github.io/palmerpenguins/
- Poisson Distribution: R Core Team (2024). Poisson Distribution Functions. R: A language and environment for statistical computing. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/Poisson
- Ross, S. M. (2023). Introduction to Probability and Statistics for Engineers and Scientists. Elsevier.