Student Score and Hypothesis Testing

This report is prepared by Khusbu Shah(ID - A20564069)

Problem Statement:

Our objective is to analyse the StudentPerformance.csv dataset to formulate relevant hypothesis and explore various scenarios that can lead to meaningful insights. These findings will support data-driven decision-making and help identify key factors influencing student performance.

We have divided this report into 3 parts-

1. Dataset Introduction and Exploration
2. Modelling and hypothesis testing
3. Conclusion

Part 1: Dataset Introduction:

StudentPerformance.csv file contains individual information about students along with their scores. Below are the list of attributes present in the file:

- Gender – gender of the student : Male, Female
- Race/Ethnicity – race of the student : group A, group B, group C, group D, group E
- Parent level of education – what is the education level of parent : bachelor's degree, some college, master's degree, associate's degree, high school, some high school
- Lunch – type of lunch : standard, free/reduced
- Test preparation- additional test preparation selected : none, completed.
- Math score – score in Math
- Reading score – score in reading
- Writing score – score in writing

Before we begin developing hypothesis, we load the dataset in R for further analysis.

Step 1: Load the dataset in R

We load the dataset in R and check for any missing or null values.

In [2]:
```
#import packages
install.packages('gridExtra')
```

```
library('dplyr')
library('ggplot2')
library('gridExtra')
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)


Attaching package: ‘gridExtra’


The following object is masked from ‘package:dplyr’:

    combine
```

In [2]:
```
#Importing data from csv file
student_data<-data.frame(read.csv('/content/sample_data/StudentsPerformance.csv'

#Reading 10 records from the file to check if the file is imported properly
head(student_data,10)
```

A data.frame: 10 × 8

	gender	race.ethnicity	parental.level.of.education	lunch	test.preparation.course
	<chr>	<chr>	<chr>	<chr>	<chr>
1	female	group B	bachelor's degree	standard	none
2	female	group C	some college	standard	completed
3	female	group B	master's degree	standard	none
4	male	group A	associate's degree	free/reduced	none
5	male	group C	some college	standard	none
6	female	group B	associate's degree	standard	none
7	female	group B	some college	standard	completed
8	male	group B	some college	free/reduced	none
9	male	group D	high school	free/reduced	completed
10	female	group B	high school	free/reduced	none

Step 2: Check for any missing or null values.

There are no missing values in the file.

In [8]:
```
#Check for null/blank values
glimpse(student_data)
colSums(is.na(student_data))
```

```
Rows: 1,000
Columns: 8
$ gender                      <chr> "female", "female", "female", "male", "mal…
$ race.ethnicity              <chr> "group B", "group C", "group B", "group A"…
$ parental.level.of.education <chr> "bachelor's degree", "some college", "mast…
$ lunch                       <chr> "standard", "standard", "standard", "free/…
$ test.preparation.course     <chr> "none", "completed", "none", "none", "none…
$ math.score                  <int> 72, 69, 90, 47, 76, 71, 88, 40, 64, 38, 58…
$ reading.score               <int> 72, 90, 95, 57, 78, 83, 95, 43, 64, 60, 54…
$ writing.score               <int> 74, 88, 93, 44, 75, 78, 92, 39, 67, 50, 52…
```

gender: 0 **race.ethnicity:** 0 **parental.level.of.education:** 0 **lunch:** 0
test.preparation.course: 0 **math.score:** 0 **reading.score:** 0 **writing.score:** 0

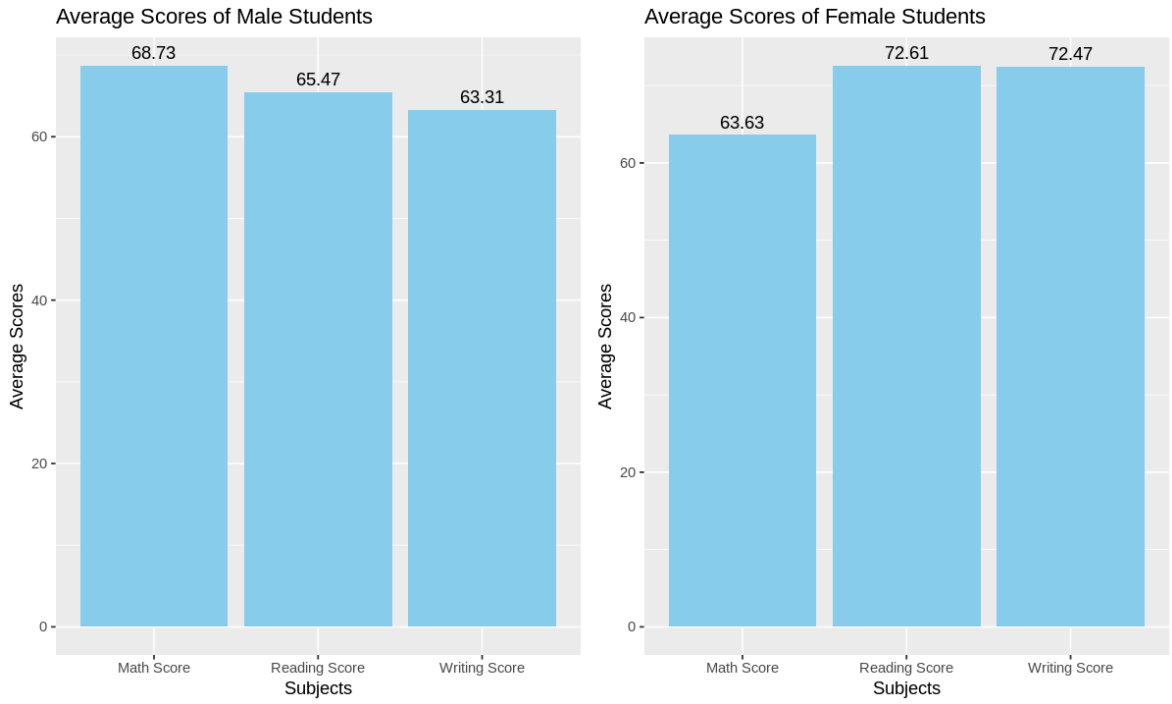
Step 3: Data Exploration:

We explore the dataset to understand the data distribution with the help of different visualization.

A. Mean scores for math, reading and writing for male and female students separately.

From the below visualization we can observe that the average score of male students in Math is greater than female students. Whereas, female students have greater average score in reading and writing compared to male students.

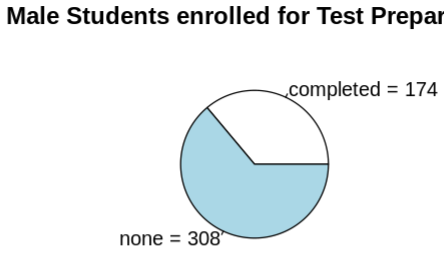
In [13]:
```
#1. Bar chart of average math, reading and writing score based on gender
male_avg_math<-mean(student_data$math.score[student_data$gender == 'male'], na.r
male_avg_read<-mean(student_data$reading.score[student_data$gender == 'male'], n
male_avg_write<-mean(student_data$writing.score[student_data$gender == 'male'],

female_avg_math<-mean(student_data$math.score[student_data$gender == 'female'],
female_avg_read<-mean(student_data$reading.score[student_data$gender == 'female'
female_avg_write<-mean(student_data$writing.score[student_data$gender == 'female

avg_scores<-data.frame(male_avg_scores=c(male_avg_math,male_avg_read,male_avg_wr
                       female_avg_scores=c(female_avg_math,female_avg_read,femal
                        subjects=c('Math Score','Reading Score','Writing Score')

plot1<-ggplot(avg_scores,aes(x=subjects,y=male_avg_scores))+
  geom_bar(stat='identity',fill='skyblue')+
  geom_text(aes(label=round(male_avg_scores,2)),vjust=-0.5,color='black')+
  labs(title='Average Scores of Male Students',x='Subjects',y='Average Scores')

plot2<-ggplot(avg_scores,aes(x=subjects,y=female_avg_scores))+
  geom_bar(stat='identity',fill='skyblue')+
  geom_text(aes(label=round(female_avg_scores,2)),vjust=-0.5,color='black')+
  labs(title='Average Scores of Female Students',x='Subjects',y='Average Scores'
options(repr.plot.width = 10, repr.plot.height = 6)
grid.arrange(plot1,plot2,ncol=2)
```

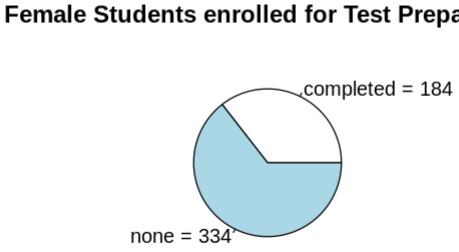


B. Count of Male and Female students completed test preparation.

From the below diagram, we can observe that count of test preparation completed by female students is higher than male students.

In [10]:
```
par(mfrow=c(2,1))
male_count<-sum(student_data$gender=='male')
female_count<-sum(student_data$gender=='female')

male_test_prep<-sum(student_data$test.preparation.course=='completed' & student_
female_test_prep<-sum(student_data$test.preparation.course=='completed' & studen

male_data<-c(male_test_prep,(male_count-male_test_prep))
female_data<-c(female_test_prep,(female_count-female_test_prep))

male_pie_labels=paste(c('completed','none'),'=',male_data)
pie(male_data,labels=male_pie_labels,main='Male Students enrolled for Test Prepa

female_pie_labels=paste(c('completed','none'),'=',female_data)
pie(female_data,labels=female_pie_labels,main='Female Students enrolled for Test
```

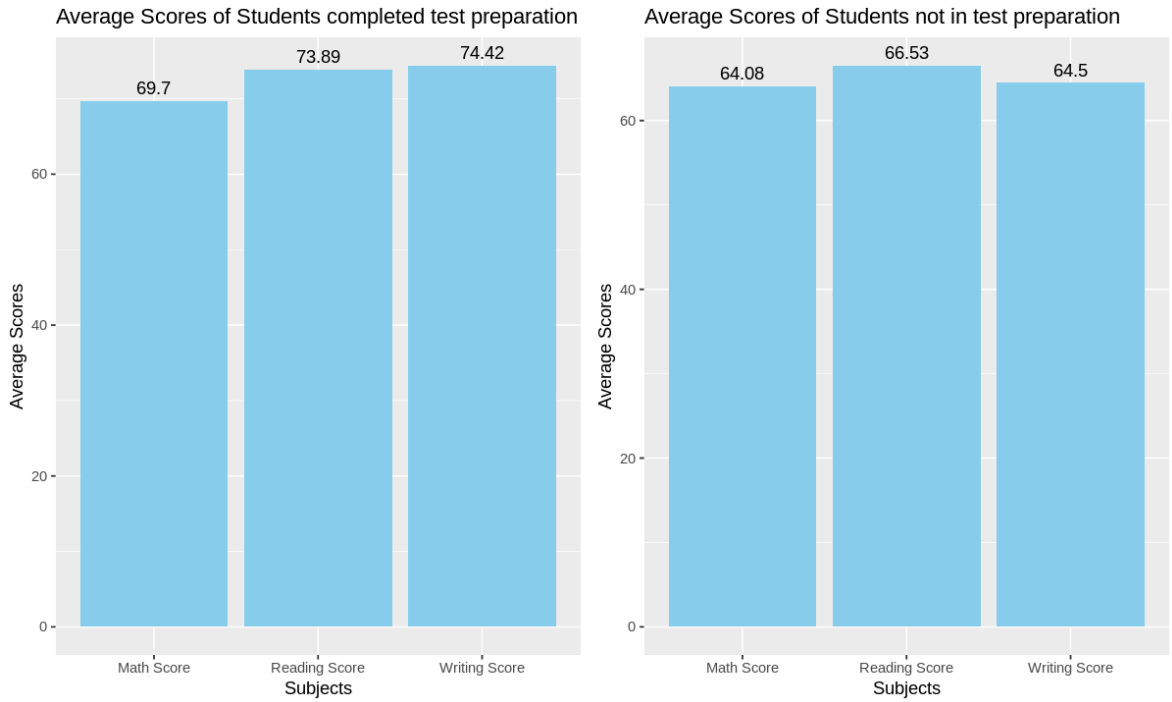




C. Comparison of average score of students who have completed test preparation vs none.

From the below diagram, we can observe that there is an improvement in the scores of students who have completed test preparation. There is improvement in the scores of Reading and Writing compared to Math.

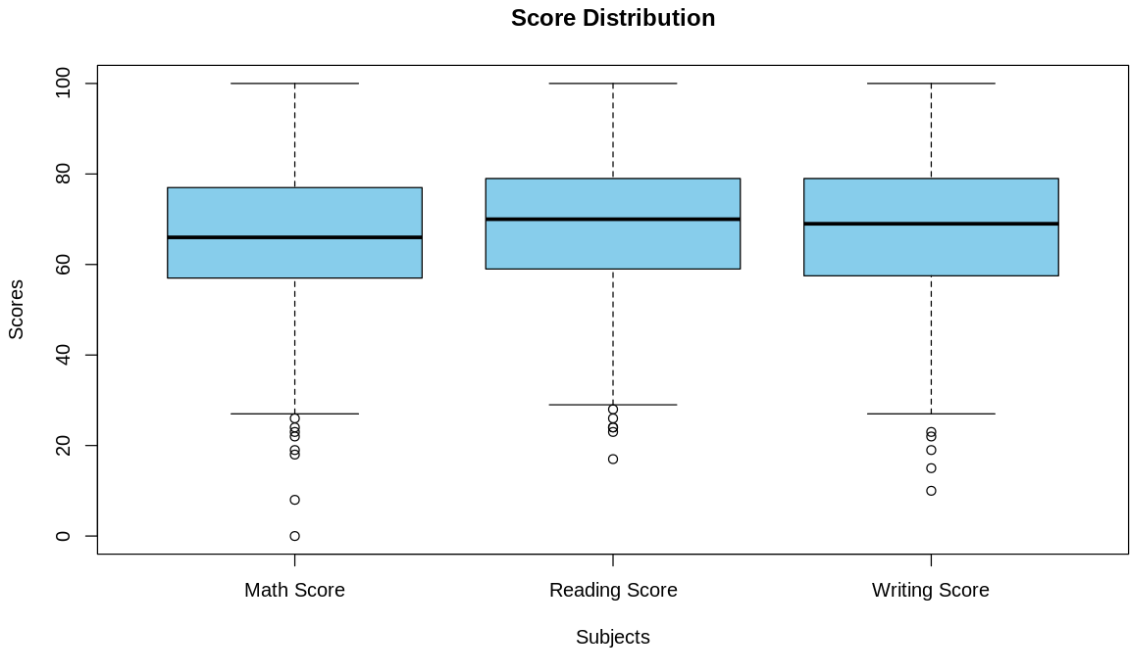
In [12]:
```
avg_math_test<-mean(student_data$math.score[student_data$test.preparation.course
avg_read_test<-mean(student_data$reading.score[student_data$test.preparation.cou
avg_write_test<-mean(student_data$writing.score[student_data$test.preparation.co

avg_math_not_test<-mean(student_data$math.score[student_data$test.preparation.co
avg_read_not_test<-mean(student_data$reading.score[student_data$test.preparation
avg_write_not_test<-mean(student_data$writing.score[student_data$test.preparatio


avg_scores_test<-data.frame(avg_scores_test=c(avg_math_test,avg_read_test,avg_wr
                            avg_score_nt=c(avg_math_not_test,avg_read_not_test,a
                       subjects=c('Math Score','Reading Score','Writing Score'))

plot3<-ggplot(avg_scores_test,aes(x=subjects,y=avg_scores_test))+
  geom_bar(stat='identity',fill='skyblue')+
  geom_text(aes(label=round(avg_scores_test,2)),vjust=-0.5,color='black')+
  labs(title='Average Scores of Students completed test preparation',x='Subjects

plot4<-ggplot(avg_scores_test,aes(x=subjects,y=avg_score_nt))+
  geom_bar(stat='identity',fill='skyblue')+
  geom_text(aes(label=round(avg_score_nt,2)),vjust=-0.5,color='black')+
  labs(title='Average Scores of Students not in test preparation',x='Subjects',y

options(repr.plot.width = 10, repr.plot.height = 6)
grid.arrange(plot3,plot4,ncol=2)
```



D. Distribution of Math, reading and writing scores

From the below box plot, we understand the score distribution for each subject.

In [15]:
```
par(mfrow=c(1,1))
boxplot(student_data$math.score,student_data$reading.score,student_data$writing.
        main='Score Distribution',xlab='Subjects',ylab='Scores',
        names=c('Math Score','Reading Score','Writing Score'),
        col = 'skyblue')
```



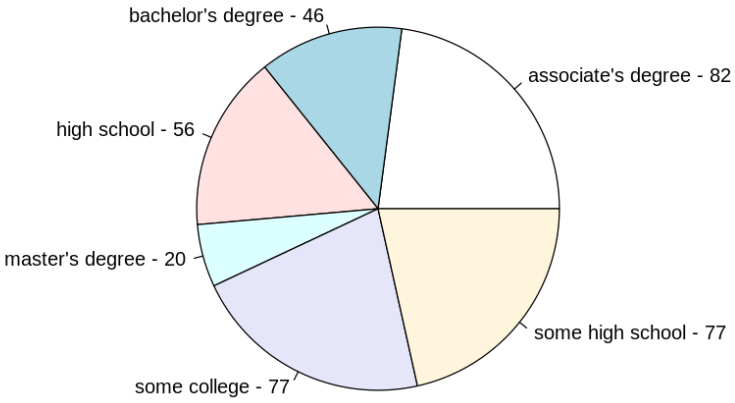
E. Distribution of Parent level of education

The below diagrams give us the count of students who prefer and not prefer test preparation based on education level of parents. This helps us in understanding which parents enrol their children for test preparation. From the diagram we understand that

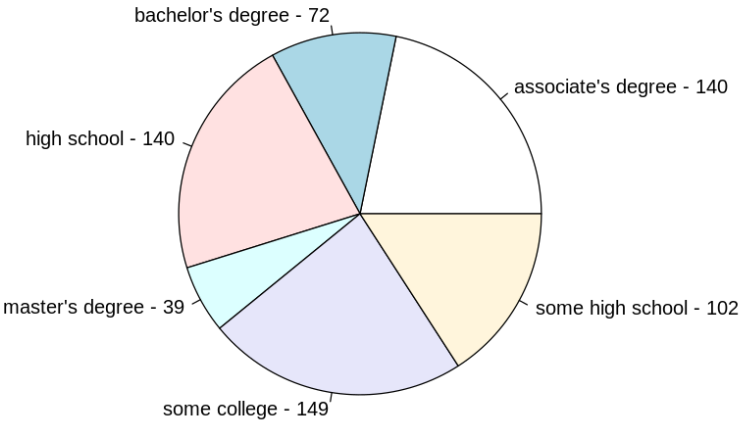
the count of students whose parents education level is - some college, some high school and associates's degree is higher to complete test preparation compared to rest others.

In [16]:
```
par(mfrow=c(1,1))
parent_data<-table(student_data$parental.level.of.education[student_data$test.pr
parent_label=paste(names(parent_data),"-",parent_data)

pie(parent_data,labels=parent_label,main='Parent level of education for students

parent_nt_data<-table(student_data$parental.level.of.education[student_data$test
parent_nt_label=paste(names(parent_nt_data),"-",parent_nt_data)

pie(parent_nt_data,labels=parent_nt_label,main='Parent level of education for st
```

Parent level of education for students completed test preparation



Parent level of education for students not in test preparation



Part 2. Modelling and Hypothesis Testing

To model and test various scenarios, we will use Bayesian analysis—a powerful approach that combines prior knowledge or assumptions with observed data. Unlike traditional methods that rely solely on the current dataset, Bayesian analysis updates our understanding as new data becomes available.

In the context of student performance, this method is especially useful for identifying uncertainty and estimating the likelihood of different outcomes. For example, we can assess how likely it is that a student will benefit from a test preparation course. This approach is particularly valuable when working with small or incomplete datasets, as it allows us to incorporate additional knowledge rather than relying only on limited observations.

By applying Bayesian models, we gain more deeper insights, leading to more informed decision-making in education planning.

Model 1: What is the probability that Male student performs better than Female student in Math, Reading and Writing

In this scenario, we want to know how well the male students are performing than female student in individual subjects. This helps in identifying students and focus on them to improve their scores.

Subject 1: Math

In this case, male student's math score is compared to female student's math score which can be seen as conditional probability. This can be written as:

P(Male Math Score > Female Math Score | Gender)

In [3]:
```
#Model 1 A: What is the probability that Male student performs better than Femal

#Score calculation
m_male_scores <- student_data$math.score[student_data$gender == "male"]
m_female_scores <- student_data$math.score[student_data$gender == "female"]

# Calculate the number of times a male student scores higher than a female stude
m_num_comparisons <- sum(outer(m_male_scores, m_female_scores, ">"))

# Calculate the total number of comparisons
m_total_comparisons <- length(m_male_scores) * length(m_female_scores)

# Calculate the probability that a male performs better than a female
m_prob_male_better_than_female <- m_num_comparisons / m_total_comparisons

# Output the result
cat("Probability for math - ",m_prob_male_better_than_female)
```

```
Probability for math -  0.5831197
```

Result for Math:

We get the probability value as 0.5831 which means that there is 58% chance that the male student will perform better than female student in Math subject.

Subject 2: Reading

Using the logic used for Math subject, we will apply similar logic to Reading subject to check the probability using R.

In [4]:
```r
#Model 1 B: What is the probability that Male student performs better than Femal

#Score calculation
r_male_scores <- student_data$reading.score[student_data$gender == "male"]
r_female_scores <- student_data$reading.score[student_data$gender == "female"]

# Calculate the number of times a male student scores higher than a female stude
r_num_comparisons <- sum(outer(r_male_scores, r_female_scores, ">"))

# Calculate the total number of comparisons
r_total_comparisons <- length(r_male_scores) * length(r_female_scores)

# Calculate the probability that a male performs better than a female
r_prob_male_better_than_female <- r_num_comparisons / r_total_comparisons

# Output the result
cat("Probability for reading - ",r_prob_male_better_than_female)
```

```
Probability for reading -  0.3479189
```

Result for Reading:

We get the probability value as 0.3479. This means that there is 34% chance that male students will perform better than female student in reading. This probability value is less which indicates that additional focus on male students is essential for better scores.

Subject 3: Writing

Using the logic used for other 2 subjects, we calculated the probability of male student performing better than female student in Writing.

In [5]:
```r
#Model 1C: What is the probability that Male student performs better than Female

#Score calculation
w_male_scores <- student_data$writing.score[student_data$gender == "male"]
w_female_scores <- student_data$writing.score[student_data$gender == "female"]

# Calculate the number of times a male student scores higher than a female stude
w_num_comparisons <- sum(outer(w_male_scores, w_female_scores, ">"))

# Calculate the total number of comparisons
w_total_comparisons <- length(w_male_scores) * length(w_female_scores)
```

```
# Calculate the probability that a male performs better than a female
w_prob_male_better_than_female <- w_num_comparisons / w_total_comparisons

# Output the result
cat("Probability for Writing - ",w_prob_male_better_than_female)
```

```
Probability for Writing -  0.3106226
```

Result for Writing:

We get the probability value as 0.3106. This means that there is 31% chance that male students will perform better than female students. Again, this value is less which indicates that additional focus is required.

Model 2: What is the probability that female student performs better than male student in Math, Reading and Writing subjects.

In this scenario, we want to know how well the female students are performing than male student in individual subjects.

Subject 1: Math

In this case, female student's math score is compared to male student's math score which is an example of conditional probability. This probability calculation can be written as:

P(Female Math Score > Male Math Score | Gender)

In [7]:
```
#Model 2A: What is the probability that Female student performs better than male

#Score calculation
fm_male_scores <- student_data$math.score[student_data$gender == "male"]
fm_female_scores <- student_data$math.score[student_data$gender == "female"]

# Calculate the number of times a male student scores higher than a female stude
fm_num_comparisons <- sum(outer(fm_female_scores, fm_male_scores, ">"))

# Calculate the total number of comparisons
fm_total_comparisons <- length(fm_male_scores) * length(fm_female_scores)

# Calculate the probability that a male performs better than a female
fm_prob_female_better_than_male <- fm_num_comparisons / fm_total_comparisons

# Output the result
cat("Probability for math - ",fm_prob_female_better_than_male)
```

```
Probability for math -  0.3983242
```

Results: The probability calculated is 0.3983, which means probability that female students are performing better than male students in Math is 39%. This value indicates that female student need more focus in Math compared to male student.

Subject 2: Reading

We use same logic as used for math subject and calculate the probability.

In [8]:
```
#Model 2 B: What is the probability that Male student performs better than Femal

#Score calculation
fr_male_scores <- student_data$reading.score[student_data$gender == "male"]
fr_female_scores <- student_data$reading.score[student_data$gender == "female"]

# Calculate the number of times a male student scores higher than a female stude
fr_num_comparisons <- sum(outer(fr_female_scores, fr_male_scores, ">"))

# Calculate the total number of comparisons
fr_total_comparisons <- length(fr_male_scores) * length(fr_female_scores)

# Calculate the probability that a male performs better than a female
fr_prob_female_better_than_male <- fr_num_comparisons / fr_total_comparisons

# Output the result
cat("Probability for Reading - ",fr_prob_female_better_than_male)
```

```
Probability for Reading -  0.6336492
```

Result:

The probability calculated is 0.6336, which means that probability that female student will perform better than male student in reading is 63%. This is a good probability which indicates females are performing better than male.

Subject 3: Writing

We use same calculation as used for math subject and calculate the probability.

In [9]:
```
#Model 2C: What is the probability that Male student performs better than Female

#Score calculation
fw_male_scores <- student_data$writing.score[student_data$gender == "male"]
fw_female_scores <- student_data$writing.score[student_data$gender == "female"]

# Calculate the number of times a male student scores higher than a female stude
fw_num_comparisons <- sum(outer(fw_female_scores, fw_male_scores, ">"))

# Calculate the total number of comparisons
fw_total_comparisons <- length(fw_male_scores) * length(fw_female_scores)

# Calculate the probability that a male performs better than a female
fw_prob_female_better_than_male <- fw_num_comparisons / fw_total_comparisons

# Output the result
cat("Probability for writing - ",fw_prob_female_better_than_male)
```

```
Probability for writing -  0.672039
```

Result:

The probability calculated is 0.6720, which means that probability that female student will perform better than male student in reading is 67%. This is a good probability which indicates females are performing better than male.

Conclusion from Model 1 and Model 2:

Model 1 and Model 2 helps us understand the performance of male and female students in Math, Writing and Reading subject by comparing their scores against each other.

- Male students are performing better than female students in Math subject. But there are chances of improvement. But Male students need additional attention in Reading and Writing compared to female students.
- Female students are performing better than male students in Reading and Writing but need to focus more on math.

Model 3: What is the probability that the student will perform better after test preparation completion.

In this model we want to understand how the test preparation is helping students in improving their score. We are calculating probability of student performance after test preparation is completed and comparing that against the normal average. Below is the calculation logic for Math. Similar logic is applied for calculating probability for Reading and Writing.

Calculation logic:

1. Calculating the priors:

P(T) = probability of the student completed test preparation. P(B) = probability of the student performs better ie score > average.

2. Calculating the likelihood:

P(T|B) = probability that the student who performs better has completed test preparation.

3. Calculating the posterior:

P(B|T) = probability that a student who has completed test preparation performs better.

In [27]:
```
#Model 3: What is the probability that the student will perform better after tes

#average math score
average_math_score <- mean(student_data$math.score, na.rm = TRUE)
average_read_score <- mean(student_data$reading.score, na.rm = TRUE)
average_write_score <- mean(student_data$writing.score, na.rm = TRUE)

# score > average score
student_data$better_performance_m <- student_data$math.score > average_math_scor
student_data$better_performance_r <- student_data$reading.score > average_read_s
student_data$better_performance_w <- student_data$writing.score > average_write_

#prior probabilities
P_T <- mean(student_data$test.preparation.course == "completed")
P_B_M <- mean(student_data$better_performance_m)
P_B_R <- mean(student_data$better_performance_r)
P_B_W <- mean(student_data$better_performance_w)

# P(T | B)
```

```
P_T_given_B_M <- mean(student_data$test.preparation.course == "completed" & stud
P_T_given_B_R <- mean(student_data$test.preparation.course == "completed" & stud
P_T_given_B_W <- mean(student_data$test.preparation.course == "completed" & stud


# P(B | T)
P_B_given_T_M <- mean(student_data$better_performance_m & student_data$test.prep
P_B_given_T_R <- mean(student_data$better_performance_r & student_data$test.prep
P_B_given_T_W <- mean(student_data$better_performance_w & student_data$test.prep

# Apply Bayes' Rule
P_B_given_T_value_m <- (P_B_given_T_M * P_T) / P_B_M
P_B_given_T_value_r <- (P_B_given_T_R * P_T) / P_B_R
P_B_given_T_value_w <- (P_B_given_T_W * P_T) / P_B_W

# result
cat("Probability of math -",P_B_given_T_value_m)
cat("\nProbability of reading -",P_B_given_T_value_r)
cat("\nProbability of writing -",P_B_given_T_value_w)
```

```
Probability of math - 0.1532211
Probability of reading - 0.1632982
Probability of writing - 0.1713086
```

Result:

The posterior probability for math = 0.1532.

The posterior probability for reading = 0.1632

The posterior probability for writing = 0.1713

These values are close to 0 which means that there is no significant increase in the scores of these students after completing the test preparation. This indicates that after completion of test preparation, there is not much significant impact in the scores of student.

Model 4: What is the probability of the student scoring above average considering the parental level of education.

In this model we will calculate the probability of the student scoring above average considering parental education level.

Different Parental Education level are - bachelor's degree , some college , master's degree, associate's degree, high school, some high school.

Calculation Logic:

1. Calculating Priors:

P(E) = Probability of students having certain parental level of education.

P(B) = Probability of the student performing above average. Here we will consider total scores (ie math score + reading score + writing score) 2. Calculating likelihood:

P(E|B) = The probability of the student having certain parental education that are performing better. 3. Calculating Posterior:

P(B|E) = The probability of the student performing better having certain parental education.

In [28]:
```
#Model 4: Probability of student performing better based on parental level of ed
total_score=student_data$math.score+student_data$reading.score+student_data$writ
avg_total_score=mean(student_data$math.score+student_data$reading.score+student_
student_data$better_performance <- total_score > avg_total_score


for (level in unique(student_data$parental.level.of.education)) {
  # P(P): Probability of better performance
  P_B <- mean(student_data$better_performance)

  # P(E): Probability of this education level
  P_E <- mean(student_data$parental.level.of.education == level)

  # P(E ∩ P): Joint probability
  P_E_and_P <- mean(student_data$parental.level.of.education == level & student_

  # P(E | P)
  P_E_given_B <- P_E_and_P / P_B

  # Bayes' Rule: P(P | E)
  P_B_given_E <- (P_E_given_B * P_B) / P_E

  # Print result
  cat("Probability of better performance given parental education is", level, ":
}
```

```
Probability of better performance given parental education is bachelor's degree :
0.6186
Probability of better performance given parental education is some college : 0.53
98
Probability of better performance given parental education is master's degree :
0.661
Probability of better performance given parental education is associate's degree
: 0.5766
Probability of better performance given parental education is high school : 0.403
1
Probability of better performance given parental education is some high school :
0.4749
```

Result:

We can observe that students perform well whose parents have Master's degree while students perform poor whose parents are educated from some college level. This indicates that parental education level is directly proportional to student's performance. If the parent is educated than the student's performance are high and viceversa.

Summary:

The below value represent what is the probability of student performing better.

bachelor's degree - 61.86%

some college - 53.98%

master's degree - 66.1%

associate's degree - 57.66%

high school - 40.31%

some high school - 47.49%

Model 5: Calculate probability of the student with features – gender - male, race - group B, test preparation course – completed and none and parental education – bachelor's degree will be a high performer.

In this model, we are considering a scenario of student with below features and calculating the posterior probability of the student being a high performer. We are considering 2 cases – test preparation completed and none to check its impact on performance.

Below are the attributes considered in this scenario:

- Gender – Male
- Race – Group B
- Test preparation course – Case 1: Completed and Case 2: none
- Parental level of education – Bachelor's degree

Calculation Logic:

In this scenario the data follows binomial distribution. Here,

θ: true probability of the student being high performer based on given features

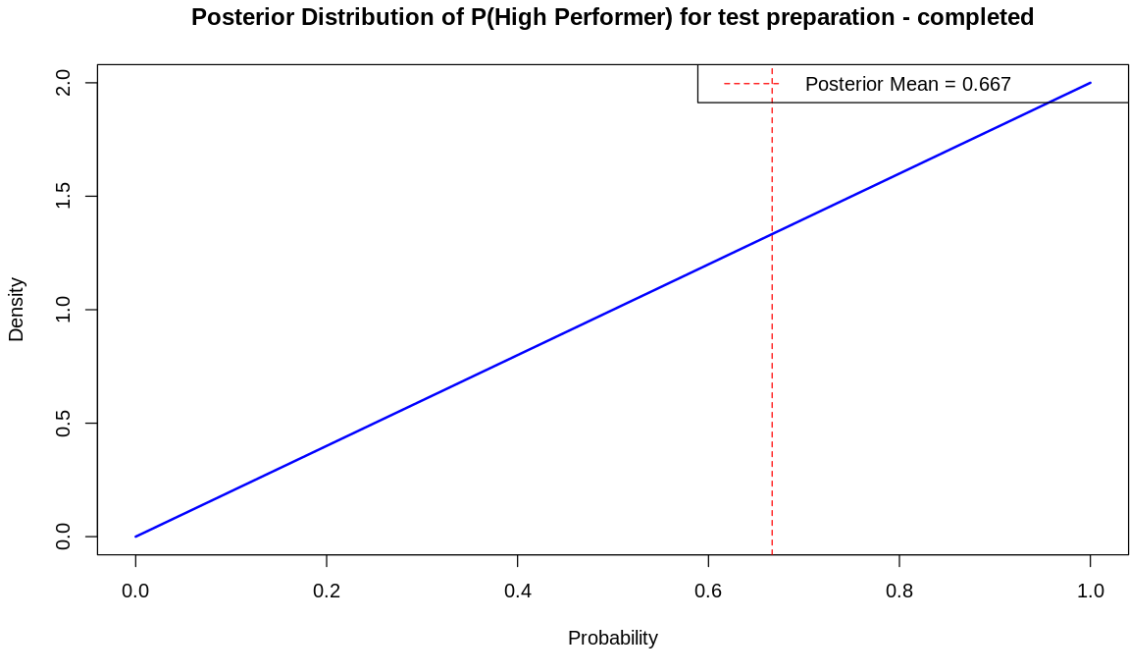
we need to calculate P(θ|data)

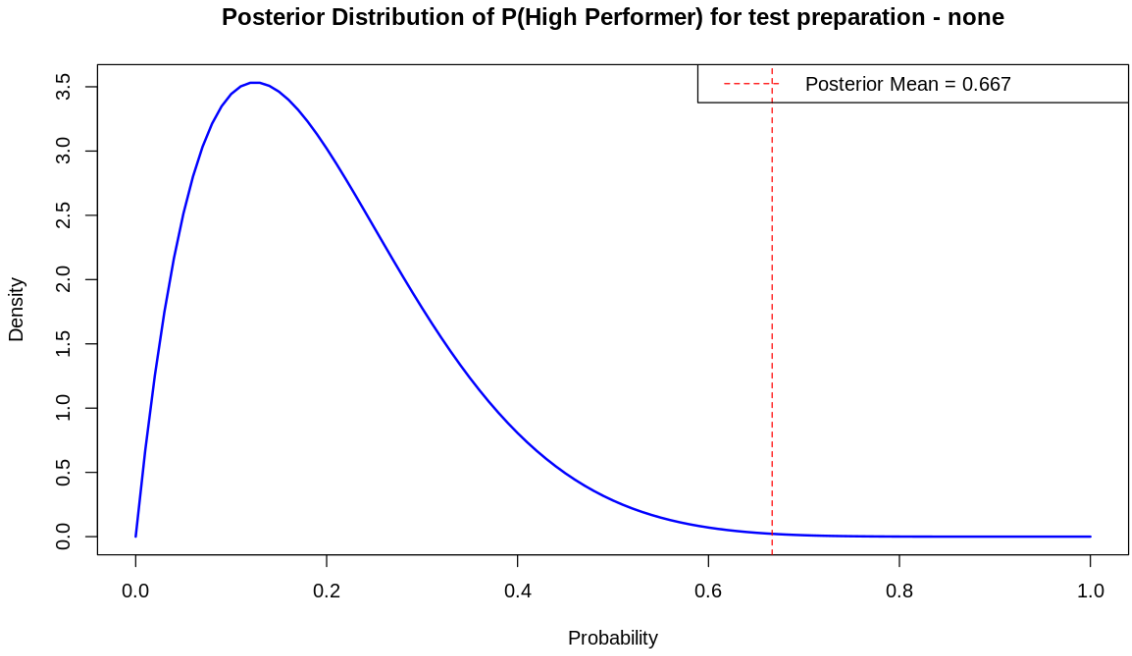
We are considering uniform prior ie prior parameters: alpha=beta=1

In [31]:
```
#Model 5: Calculate posterior probability of the student being high performer ba

#Calculate high performer
student_data$high_performer <- ifelse(total_score > avg_total_score, 1, 0)

#Filter data
filter_data <- subset(student_data,
                      gender == "male" &
                        race.ethnicity == "group B" &
                        parental.level.of.education == "bachelor's degree" &
                        test.preparation.course == "completed")

k <- sum(filter_data$high_performer)  # number of high performers
n <- nrow(filter_data)    # total students in this group

#Choose prior
alpha <- 1
beta <- 1
```

```r
# Posterior distribution
post_alpha <- alpha + k
post_beta <- beta + n - k

# Posterior estimate of theta (mean of Beta)
posterior_mean <- post_alpha / (post_alpha + post_beta)

cat("Posterior probability of high performance when test preparation is complete

#Curve visualization
curve(dbeta(x, post_alpha, post_beta), from=0, to=1,
      main="Posterior Distribution of P(High Performer) for test preparation - c
      ylab="Density", xlab="Probability", col="blue", lwd=2)
abline(v = posterior_mean, col = "red", lty = 2)
legend("topright", legend=paste("Posterior Mean =", round(posterior_mean, 3)), c

#Filter data
filter_data1 <- subset(student_data,
                       gender == "male" &
                         race.ethnicity == "group B" &
                         parental.level.of.education == "bachelor's degree" &
                         test.preparation.course == "none")

k1 <- sum(filter_data1$high_performer)  # number of high performers
n1 <- nrow(filter_data1)    # total students in this group

# Posterior distribution
post_alpha1 <- alpha + k1
post_beta1 <- beta + n1 - k1

# Posterior estimate of theta (mean of Beta)
posterior_mean1 <- post_alpha1 / (post_alpha1 + post_beta1)

cat("Posterior probability of high performance when test preparation is none:",

#Curve visualization
curve(dbeta(x, post_alpha1, post_beta1), from=0, to=1,
      main="Posterior Distribution of P(High Performer) for test preparation - n
      ylab="Density", xlab="Probability", col="blue", lwd=2)
abline(v = posterior_mean, col = "red", lty = 2)
legend("topright", legend=paste("Posterior Mean =", round(posterior_mean, 3)), c
```

```
Posterior probability of high performance when test preparation is completed: 0.6
667
Posterior probability of high performance when test preparation is none: 0.2
```





Result:

Case 1: Gender- Male, Race- group B, Test preparation – completed, parent level of education – Bachelor's degree

Considering above features, the posterior probability of male student will be a high performer is 0.66. This means that there is 66% chance that a male student with above set of features will be a high performer.

Case 2: Gender- Male, Race- group B, Test preparation – none, parent level of education – Bachelor's degree **bold text**

Considering above features, the posterior probability of male student will be a high performer is 0.2. This means that there is 20% chance that a male student with above set of features will be a high performer.

Model Conclusion:

The results clearly indicate that the performance of the student has significantly improved after the test preparation completion considering the given set of background.

Model 6: Estimating posterior distribution of mean and variance of math, reading and writing scores of male and female students.

Instead of calculating point estimate average scores for math, reading and writing, we want know possible distribution or range of values. We will perform calculation using Bayesian normal inverse gamma model. Here,

- There is normal likelihood as the scores are continuous.
- Conjugate priors – both prior and posterior distribution are normal-inverse gamma distributed.

We are considering below value of hyperparameters for calculation:

Prior mean = 70

Prior strength = 1

Alpha0 = 2

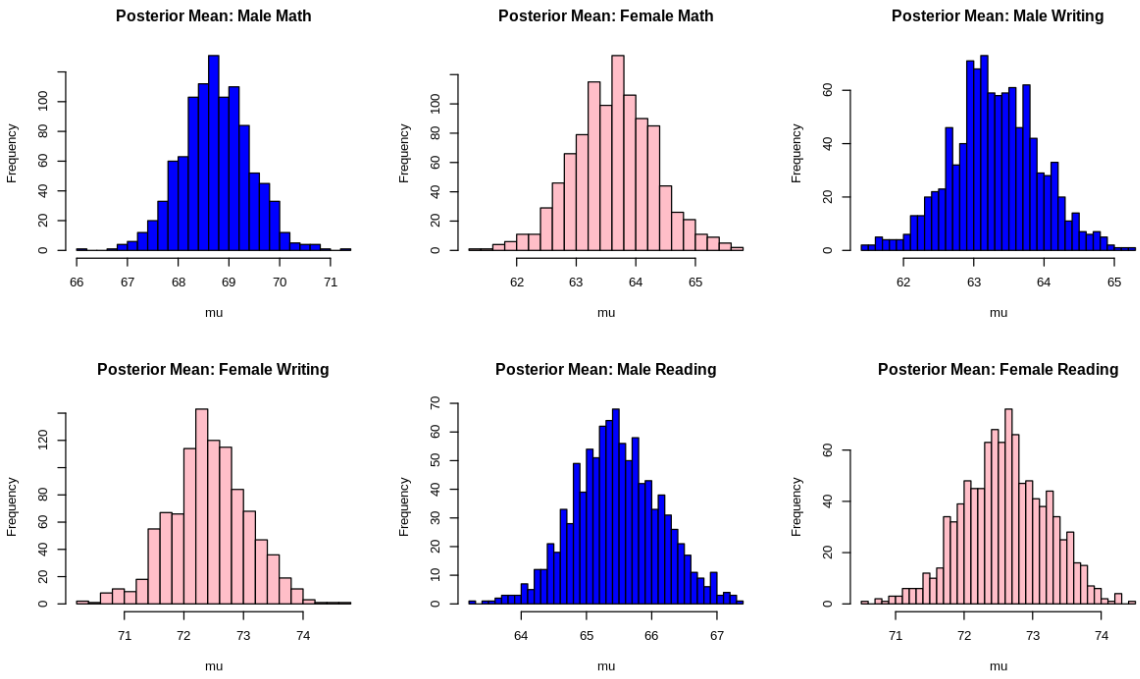
Beta0 = 50

In [32]:
```
#Model 6: Estimate posterior distribution of mean and variance of math, reading

#compute posterior parameters
compute_posterior <- function(y, mu0, kappa0, alpha0, beta0) {
  n <- length(y)
  y_bar <- mean(y)
  s2 <- var(y)

  kappa_n <- kappa0 + n
  mu_n <- (kappa0 * mu0 + n * y_bar) / kappa_n
  alpha_n <- alpha0 + n / 2
  beta_n <- beta0 + 0.5 * sum((y - y_bar)^2) +
    (kappa0 * n * (y_bar - mu0)^2) / (2 * (kappa0 + n))

  return(list(mu_n = mu_n, kappa_n = kappa_n,
              alpha_n = alpha_n, beta_n = beta_n))
}

# Simulate posterior samples
simulate_posterior <- function(posterior, n_draws = 1000) {
  sigma2_draws <- 1 / rgamma(n_draws, shape = posterior$alpha_n, rate = posterio
  mu_draws <- rnorm(n_draws, mean = posterior$mu_n, sd = sqrt(sigma2_draws / pos
  return(data.frame(mu = mu_draws, sigma2 = sigma2_draws))
}

#Split data by gender
male_math_scores <- student_data$math.score[student_data$gender == "male"]
female_math_scores <- student_data$math.score[student_data$gender == "female"]
```

```
male_writing_scores <- student_data$writing.score[student_data$gender == "male"]
female_writing_scores <- student_data$writing.score[student_data$gender == "fema

male_reading_scores <- student_data$reading.score[student_data$gender == "male"]
female_reading_scores <- student_data$reading.score[student_data$gender == "fema

#Define prior hyperparameters
mu0 <- 70
kappa0 <- 1
alpha0 <- 2
beta0 <- 50

#Compute posterior distribution
posterior_male_math <- compute_posterior(male_math_scores, mu0, kappa0, alpha0,
posterior_female_math <- compute_posterior(female_math_scores, mu0, kappa0, alph

posterior_male_writing <- compute_posterior(male_writing_scores, mu0, kappa0, al
posterior_female_writing <- compute_posterior(female_writing_scores, mu0, kappa0

posterior_male_reading <- compute_posterior(male_reading_scores, mu0, kappa0, al
posterior_female_reading <- compute_posterior(female_reading_scores, mu0, kappa0

# Simulate from the posterior distribution
samples_male_math <- simulate_posterior(posterior_male_math)
samples_female_math <- simulate_posterior(posterior_female_math)

samples_male_writing <- simulate_posterior(posterior_male_writing)
samples_female_writing <- simulate_posterior(posterior_female_writing)

samples_male_reading <- simulate_posterior(posterior_male_reading)
samples_female_reading <- simulate_posterior(posterior_female_reading)

# Results
cat("Posterior for MALES (Math):\n")
summary(samples_male_math)

cat("\nPosterior for FEMALES (Math):\n")
summary(samples_female_math)

cat("\nPosterior for MALES (Writing):\n")
summary(samples_male_writing)

cat("\nPosterior for FEMALES (Writing):\n")
summary(samples_female_writing)

cat("\nPosterior for MALES (Reading):\n")
summary(samples_male_reading)

cat("\nPosterior for FEMALES (Reading):\n")
summary(samples_female_reading)

#Visualization

par(mfrow=c(2, 3)) # Set up grid for plots

# Math Scores
hist(samples_male_math$mu, col='blue', breaks=30, main='Posterior Mean: Male Mat
hist(samples_female_math$mu, col='pink', breaks=30, main='Posterior Mean: Female

# Writing Scores
```

```
hist(samples_male_writing$mu, col='blue', breaks=30, main='Posterior Mean: Male
hist(samples_female_writing$mu, col='pink', breaks=30, main='Posterior Mean: Fem

# Reading Scores
hist(samples_male_reading$mu, col='blue', breaks=30, main='Posterior Mean: Male
hist(samples_female_reading$mu, col='pink', breaks=30, main='Posterior Mean: Fem
```

```
Posterior for MALES (Math):
       mu            sigma2      
 Min.   :66.13   Min.   :163.6  
 1st Qu.:68.29   1st Qu.:195.7  
 Median :68.71   Median :204.0  
 Mean   :68.74   Mean   :204.7  
 3rd Qu.:69.17   3rd Qu.:213.1  
 Max.   :71.21   Max.   :265.7  
Posterior for FEMALES (Math):
       mu            sigma2      
 Min.   :61.31   Min.   :193.5  
 1st Qu.:63.19   1st Qu.:228.4  
 Median :63.65   Median :238.0  
 Mean   :63.64   Mean   :238.9  
 3rd Qu.:64.10   3rd Qu.:248.4  
 Max.   :65.74   Max.   :295.1  
Posterior for MALES (Writing):
       mu            sigma2      
 Min.   :61.50   Min.   :161.0  
 1st Qu.:62.92   1st Qu.:189.4  
 Median :63.29   Median :198.1  
 Mean   :63.31   Mean   :198.4  
 3rd Qu.:63.72   3rd Qu.:206.2  
 Max.   :65.20   Max.   :245.6  
Posterior for FEMALES (Writing):
       mu            sigma2      
 Min.   :70.23   Min.   :169.7  
 1st Qu.:72.03   1st Qu.:210.0  
 Median :72.41   Median :218.5  
 Mean   :72.43   Mean   :219.3  
 3rd Qu.:72.85   3rd Qu.:228.0  
 Max.   :74.71   Max.   :263.5  
Posterior for MALES (Reading):
       mu            sigma2      
 Min.   :63.20   Min.   :163.3  
 1st Qu.:65.02   1st Qu.:184.5  
 Median :65.45   Median :192.9  
 Mean   :65.47   Mean   :192.8  
 3rd Qu.:65.91   3rd Qu.:200.2  
 Max.   :67.30   Max.   :230.7  
Posterior for FEMALES (Reading):
       mu            sigma2      
 Min.   :70.52   Min.   :176.3  
 1st Qu.:72.17   1st Qu.:197.6  
 Median :72.60   Median :205.2  
 Mean   :72.60   Mean   :206.1  
 3rd Qu.:73.03   3rd Qu.:214.7  
 Max.   :74.47   Max.   :247.9  
```



Results:

The above results and visualization gives posterior range of mean scores for reading, writing and math for male and female students.

For Male Student:

Math - 66.33 - 70.81

Reading - 63.28 - 67.25

Writing - 61.36 - 65.6

For Female Students:

Math - 61.14 - 65.82

Reading - 70.34 - 74.36

Writing - 70.35 - 74.42

Model 7: Estimate the number of students whose total score (math + reading + writing) exceeds the threshold of 180 out of 300.

We need to estimate number of students whose total score is above the threshold of 180. For this we will be considering poisson model with gamma prior. Here,

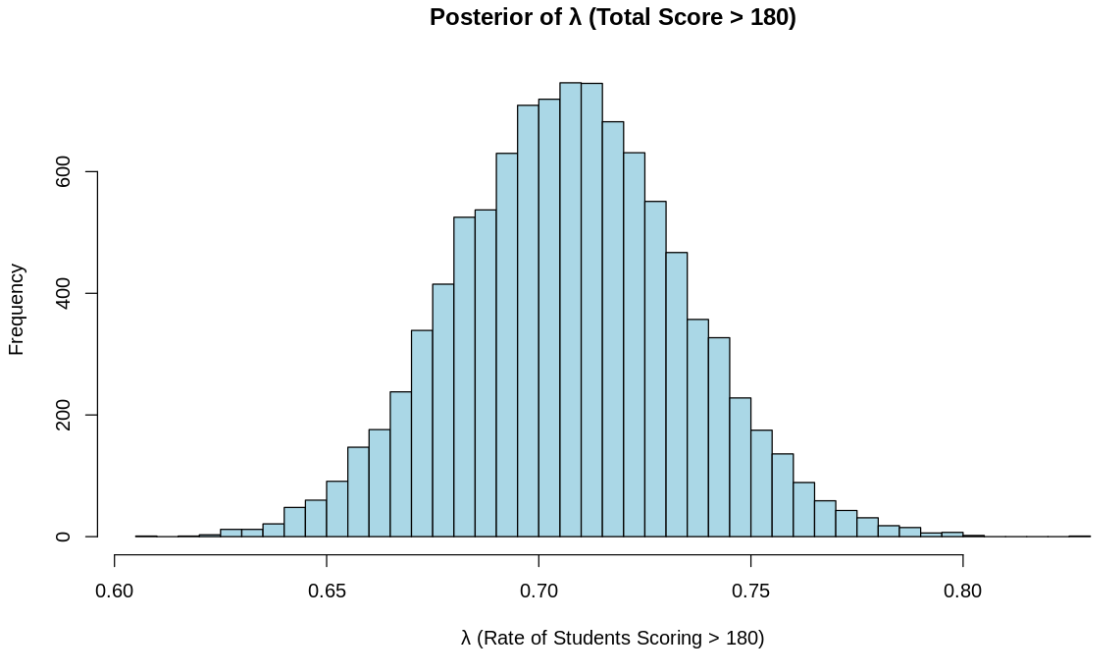
- Likelihood is Poisson distribution with rate λ
- Prior and posterior are conjugate prior with gamma distribution.

Posterior calculation:

λ|data~Gamma(α+k,β+n) where n= number of students and k=student score>180.

In [33]:
```r
#Model 7: Estimating the number of students scoring above 180.

#Calculate total score
student_data$total_score <- student_data$math.score +
  student_data$reading.score +
  student_data$writing.score

# Count how many students scored above 180
k <- sum(student_data$total_score > 180)
n <- nrow(student_data)

# Set prior parameters
alpha_prior <- 1
beta_prior <- 1

#posterior parameters
alpha_post <- alpha_prior + k
beta_post <- beta_prior + n

# Gamma distribution
set.seed(42)
posterior_lambda <- rgamma(10000, shape = alpha_post, rate = beta_post)

# Summarize posterior
posterior_mean <- mean(posterior_lambda)
cred_interval <- quantile(posterior_lambda, c(0.025, 0.975))

#results
cat("Posterior mean of lambda (rate of total score > 150):", round(posterior_mea
cat("95% credible interval:", round(cred_interval[1], 4), "-", round(cred_interv

#Visulization
par(mfrow=c(1,1))
hist(posterior_lambda, breaks = 40, col = 'lightblue',
     main = 'Posterior of λ (Total Score > 180)',
     xlab = 'λ (Rate of Students Scoring > 180)')
```

```
Posterior mean of lambda (rate of total score > 150): 0.7073
95% credible interval: 0.6551 - 0.761
```



Result:

The posterior mean value is 0.70. This means that we expect 70% of the student to score over the threshold of 180.

Credibility Interval - 95% credible interval: 0.6551 - 0.761

This means that we are 95% confident that true rate lies between 65% to 76%. This helps in understanding the performance level of student above the threshold score.

Model 8: What is the probability of the student scoring above average considering their race/ethnicity.

In this model we are calculating the probability of student scoring above average for different race/ethnicity.

Different Race/Ethnicity – group A, group B, group C, group D, group E

Calculation Logic:

1. Calculating Priors:

P(R) = Probability of students of certain race/ethnicity.

P(B) = Probability of the student performing above average. Here we will consider total scores (ie math score + reading score + writing score)

2. Calculating likelihood:

P(R|B) = The probability of the student of certain race/ethnicity that are performing better.

3. Calculating Posterior:

P(B|E) = The probability of the student performing better of certain race/ethnicity

In [34]:
```
#Model 8: What is the probability of the student scoring above average consideri

for (level in unique(student_data$race.ethnicity)) {
  # P(P): Probability of better performance
  P_B <- mean(student_data$better_performance)

  # P(E): Probability of this education level
  P_R <- mean(student_data$race.ethnicity == level)


  P_B_and_R <- mean(student_data$race.ethnicity == level & student_data$better_p

  # P(E | P)
  P_R_given_B <- P_B_and_R / P_B

  # Bayes' Rule: P(P | E)
  P_B_given_E <- (P_R_given_B * P_B) / P_R

  # Print result
  cat("Probability of better performance of certain race/ethnicity is", level, "
}
```

```
Probability of better performance of certain race/ethnicity is group B : 0.4263
Probability of better performance of certain race/ethnicity is group C : 0.5204
Probability of better performance of certain race/ethnicity is group A : 0.3483
Probability of better performance of certain race/ethnicity is group D : 0.5802
Probability of better performance of certain race/ethnicity is group E : 0.6857
```

Result:

The above result contains probability percentage of student performance based on race/ethnicity. As we can observe in the result, the chance of students performing above average who belong to group E is 68% which is a highest probability amongst all groups while the student belonging to Group A has 34.83% which is the lowest. This statistic gives an understanding of how performance of the student is related to their race/ethnicity. This helps in identifying students who need more attention.

Model 9: Classify a new student with below attributes to be high or low performer.

Consider a new student with below attributes. We need to identify if the below student will be a high performer or low.

- Gender - Female
- Race/Ethnicity – Group E
- Parent level of education – Bachelor's degree

For this model, we will consider naives bayes classification model in which we are assuming that all the attribute are independent of each other.

Calculation Logic:

- We are considering total score and threshold considered for classification is 200. Ie if the total_score > 200 then the student is a high performer else low.

- We first calculate P(high|Female, group E |bachelor's degree)
- We then calculate P(low|Female, group E |bachelor's degree)
- We then compare their results for classification.

In [35]:
```r
#Model 9: Classify new student based on certain attributes will be a high/low pe

# Define performance
student_data$performance <- ifelse(student_data$total_score > 200, "high", "low"

gender='female'
race='group E'
edu="bachelor's degree"

#prior probabilities
prior_high <- mean(student_data$performance == "high")
prior_low <- 1 - prior_high

# Likelihoods
lik_female_high <- mean(student_data$gender == gender & student_data$performance
lik_female_low <- mean(student_data$gender == gender & student_data$performance

# P(race = group E | performance)
lik_raceE_high <- mean(student_data$race.ethnicity == race & student_data$perfor
lik_raceE_low <- mean(student_data$race.ethnicity == race & student_data$perform

# P(parental education = bachelor's degree | performance)
lik_edu_high <- mean(student_data$parental.level.of.education == edu & student_d
lik_edu_low <- mean(student_data$parental.level.of.education == edu & student_da

#Compute unnormalized posteriors
posterior_high <- lik_female_high * lik_raceE_high * lik_edu_high * prior_high
posterior_low  <- lik_female_low * lik_raceE_low * lik_edu_low * prior_low

#Normalize
total_posterior <- posterior_high + posterior_low
prob_high <- posterior_high / total_posterior
prob_low <- posterior_low / total_posterior

#result
cat("P(High Performer):", round(prob_high, 4), "\n")
cat("P(Low Performer):", round(prob_low, 4), "\n")

if (prob_high > prob_low) {
  cat("Prediction: High Performer\n")
} else {
  cat("Prediction: Low Performer\n")
}
```

```
P(High Performer): 0.8176
P(Low Performer): 0.1824
Prediction: High Performer
```

Results:

From the calculations, we get the probability that the new female student with race – group E and parent level of education – bachelor's degree to be high performer is 0.8176. This means, there is 81.76% chance that the new student will be a high performer ie who will score > 200.

Part 3: Conclusion

We have taken help of bayes model in various scenario to help us understand the student's performance based on parental education level, gender and race/ethnicity, classifying students to be high or low performers, understanding test preparation benefits etc. Below is the summary of the model implemented and insights we gain from them.

Model 1 and Model 2: Gender Based Performance – Comparing male and female students' performance in individual subjects – Math, Reading and Writing.

Through this we get an understanding of who performs better in what subject and who needs more attention. Male students are performing better than female student in Math by 58%. But female students are performing better then male students in Reading and Writing by 63% and 67% respt. This means that male students need more attention in reading and writing and female students need attention in math.

Model 3: Probability of student performing better after completion of test preparation course.

Through this model, we are trying to understand if test preparation course is assisting students to perform better. The calculated probabilities are math – 0.15, reading – 0.16 and writing – 0.17. These values indicate that test preparation is not assisting well enough in score improvement. Some enhancement is required in test preparation courses to assist students in improving their scores.

Model 4: Parental level of education – probability of the student performing greater than average based on parental level of education.

Through this model we came to know that parent's level of education directly affects student's performance. Student's whose parents are educated are performing well compared to other students whose parents are not that educated.

The below value represent by what probability the student will performing better for different parental education level.

bachelor's degree - 61.86%

some college - 53.98%

master's degree - 66.1%

associate's degree - 57.66%

high school - 40.31%

some high school - 47.49%

Model 5: Calculate probability of the student with features – gender - male, race - group B, test preparation course – completed and none and parental education – bachelor's degree will be a high performer.

Here we have considered a scenario where we want to assess the probability of the student being higher performer based on certain condition. We observe that keeping gender, race and parental education same and only changing test preparation course – none/completed, we observe a significant change in probability. Students who complete test preparation course have 66% chances of being high performer whereas who have not part of test preparation course have 20% chance of being high performer.

Model 6: Estimating posterior distribution of mean and variance of math, reading and writing scores for male and female students.

Through this model we get a distribution of average scores in Math, Reading and Writing for male and female students. Rather than a point value, it provides us a range of average scores.

Model 7: Estimate the number of students whose total score (math + reading + writing) exceeds the threshold of 180 out of 300.

This model helps us in estimating the count of students being high performer ie their total score > 180. We get the results as 70% of the students score over the threshold of 180 with credible interval of 0.6551 - 0.761. this means that we are 95% confident that this value lies between 65% to 76%.

Model 8: Race/Ethnicity classification - probability of the student scoring above average considering their race/ethnicity.

Below table contains probability of student performance based on their race/ethnicity. Group E students show high performance while Group A students perform low. This helps in identifying students who need test preparation course or extra guidance for better performance.

Race/ethnicity Probability percentage

Group A - 34.83%

Group B - 63%

Group C - 52.04%

Group D - 58.02%

Group E - 68.57%

Model 9: Classify a new student with below attributes to be high or low performer.

This model helps to understand a new student's performance based on below attribute value: Gender – female, race/ethnicity – group E and parent level of education – bachelor's degree. Based on these conditions, the student turns out to be a high performer. This kind of insight helps in assigning the student the right class. Suppose if the student is low performer, then in such cases recommending extra classes or guidance is crucial.

Model Suitability and Evaluation:

These models were developed to evaluate student performance from multiple perspectives, including gender, parental education, and race/ethnicity, in order to better understand how these factors influence academic growth. Additionally, we examined the impact of test preparation courses to assess whether they provide any measurable benefit to students.

While the models offer meaningful insights, it is important to note that the limited sample size constrains the ability to draw definitive conclusions or validate the accuracy of the results. Nevertheless, the findings can still serve as a valuable foundation for informed decision-making and identifying areas for further investigation.

Conclusion:

There are several models implemented using Bayesian analysis to give various insights. These models provided a multi-dimensional evaluation, allowing for a deeper understanding of various factors influencing academic outcomes. The insights derived from these models support more informed and effective decision-making aimed at enhancing student achievement.