STAT 1150 - Assignment 1

Mykola Chudak (8043157)

2025-05-18

Instructions

To properly view the assignment questions, knit this file to .PDF and view the output.

To enter the R-based questions, add code as needed into the R code chunks given below the question, and, where applicable, replace the "Delete this text; . . . " with your own text response. Be sure when adding in text responses to never copy and paste symbols from outside of the document. Only use the symbols on your keyboard. **Do not** delete the question text, or modify any other part of the code except for the "author" in Line 3.

You will have a link in your email that takes you to the Crowdmark submission page. Once you have completed the assignment, knit it to .PDF and upload your output to Crowdmark. Make sure you set your Name and Student Number in the Author section of this document (Line 3). Do not alter the title or the date. Please note that if you do not submit a knit .PDF file, you will be given a grade of zero on the assignment.

After you knit your assignment to PDF, check your code chunks. If your code at any point runs off the page, find the nearest comma, click to the right of it, and press Enter (or Return if you are on a Mac). This will force a break in the code so that it goes onto the next line. All of your code must be readable in the final submission.

It is good practice to knit the document after each new line of code. This will enable you to easily locate any errors reported by R.

For the R-based questions, all calculations and output must be visible in the final knit PDF. Your work should be done using the same formatting, functions, and packages as in your labs and course notes, unless otherwise specified. You may speak to your classmates about ideas and what functions/arguments you may need to use but you may not directly show your code/output to your classmates.

Your full submission is due by 11:59 p.m. on Thursday, May 22. For each day that this submission is late, there will be a 25% penalty applied after the assignment is graded. i.e., there will be a 25% penalty for any submissions received on May 23, 50% on May 24, 75% on May 25, and 100% thereafter.

If you have an issue that you can't resolve without someone looking at your work (e.g., you get an error when knitting your document and you can't figure out what the error is), please see a TA in the Help Centre in 107 Allen.

Setup [0 marks]

The NBA data set contains information on 393 National Basketball Association (NBA) players for 30 different variables. Import the data set below. Make sure the data set is named NBA when you import it, **and that "Yes" is selected beside "Heading"**.

```r
NBA <- read.csv("NBA.csv")
```

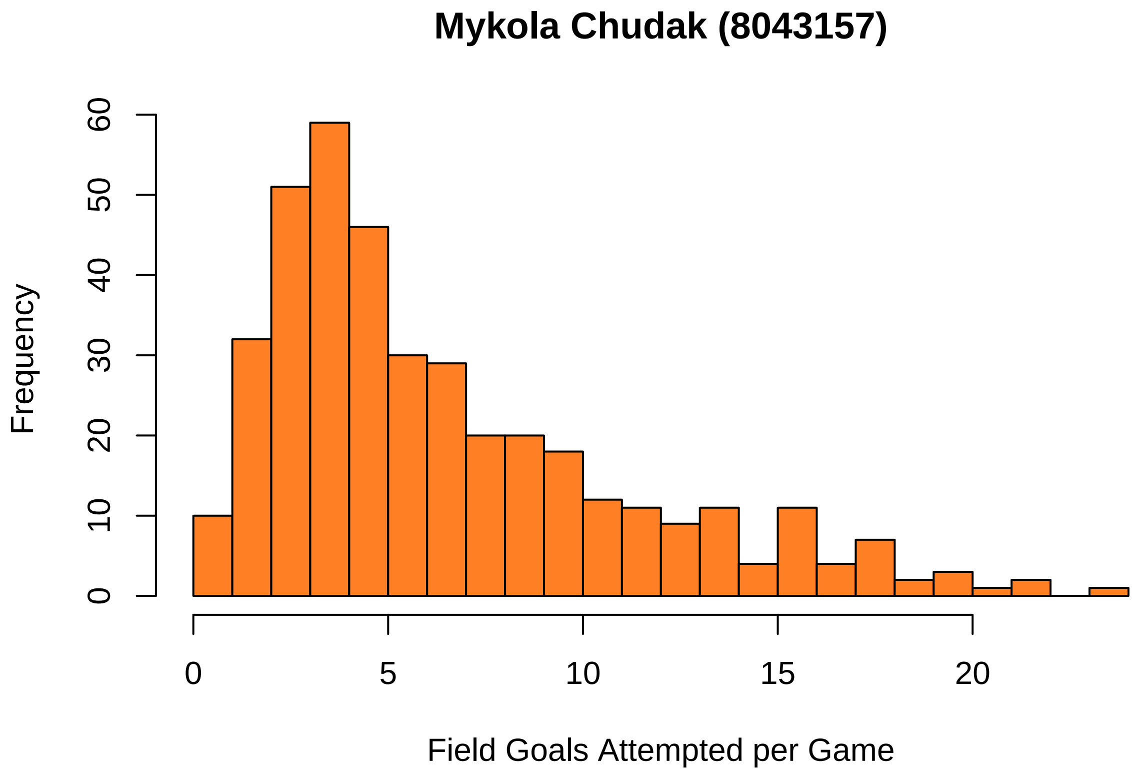
Questions [65 marks]

Question 1 [28 marks]

Part (a) [2 marks]

The variable FGA measures the average number of field goals attempted per game. Below, make a histogram of the FGA variable. Set the main title as your name and student number, set the breaks argument to 20, the x-axis label to "Field Goals Attempted per Game", and change the color to chocolate1.

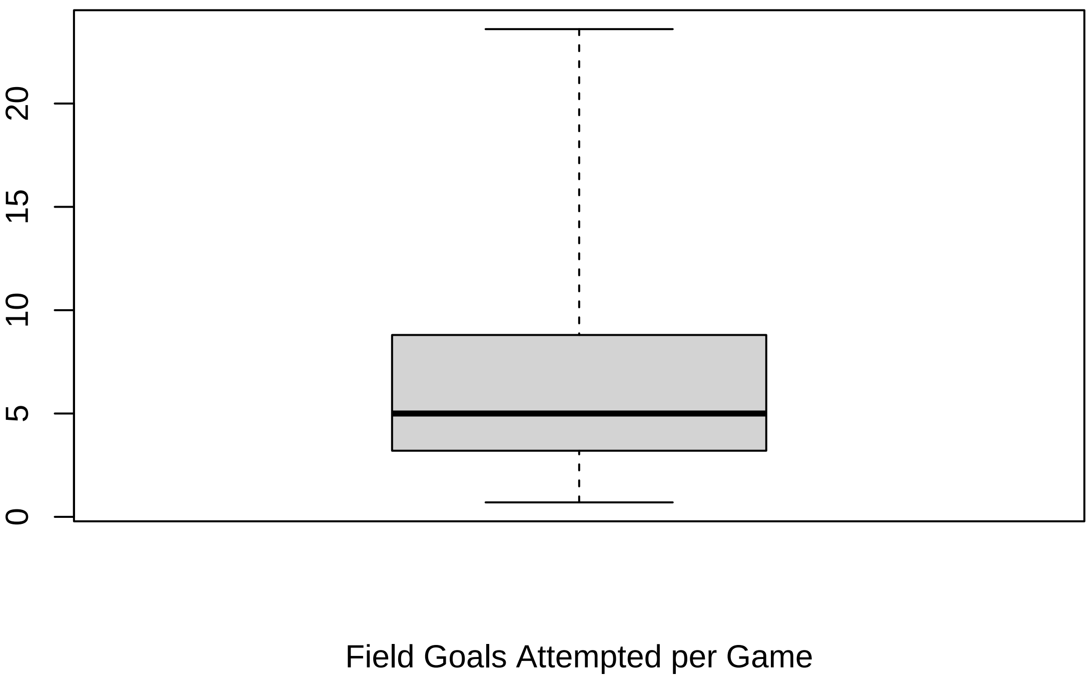
```r
hist(NBA$FGA, main = "Mykola Chudak (8043157)", breaks = 20,
     xlab = "Field Goals Attempted per Game", col = "chocolate1")
```



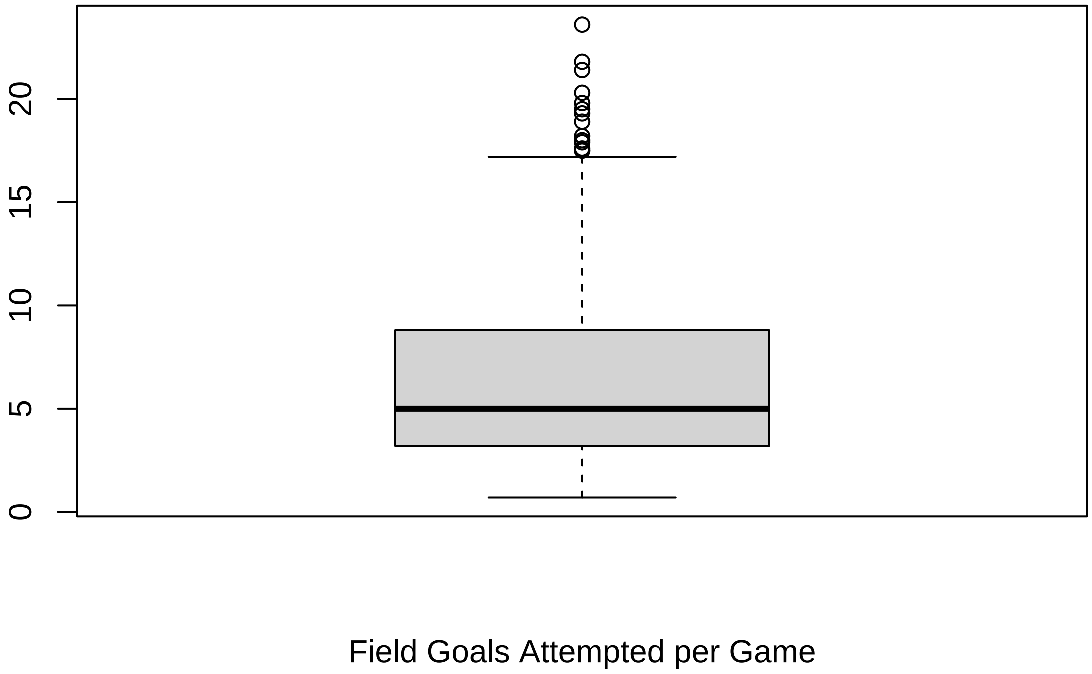
Part (b) [2 marks]

Make two horizontal boxplots of the `FGA` variable (One quantile boxplot and one outlier boxplot, in that order). For each boxplot, set the x-axis label to "Field Goals Attempted per Game".

```
boxplot(NBA$FGA, range = 0, xlab = "Field Goals Attempted per Game")
```



```
boxplot(NBA$FGA, xlab = "Field Goals Attempted per Game")
```



Part (c) [1 mark]

Based on the histogram and the boxplots, what is the shape of the distribution of the FGA variable, excluding outliers?

Skew to the right/Positively skewed

Part (d) [3 marks]

A player's average number of field goals attempted per game will be considered an outlier if it is below or above what values? To answer this question, first use R to calculate the five-number summary, and then use R as a calculator to find the requested values.

```
fivenum(NBA$FGA)
```

```
## [1]  0.7  3.2  5.0  8.8 23.6
```

```
IQR <- fivenum(NBA$FGA)[4] - fivenum(NBA$FGA)[2]
fivenum(NBA$FGA)[4] + 1.5 * IQR # Upper Fence
```

```
## [1] 17.2
```

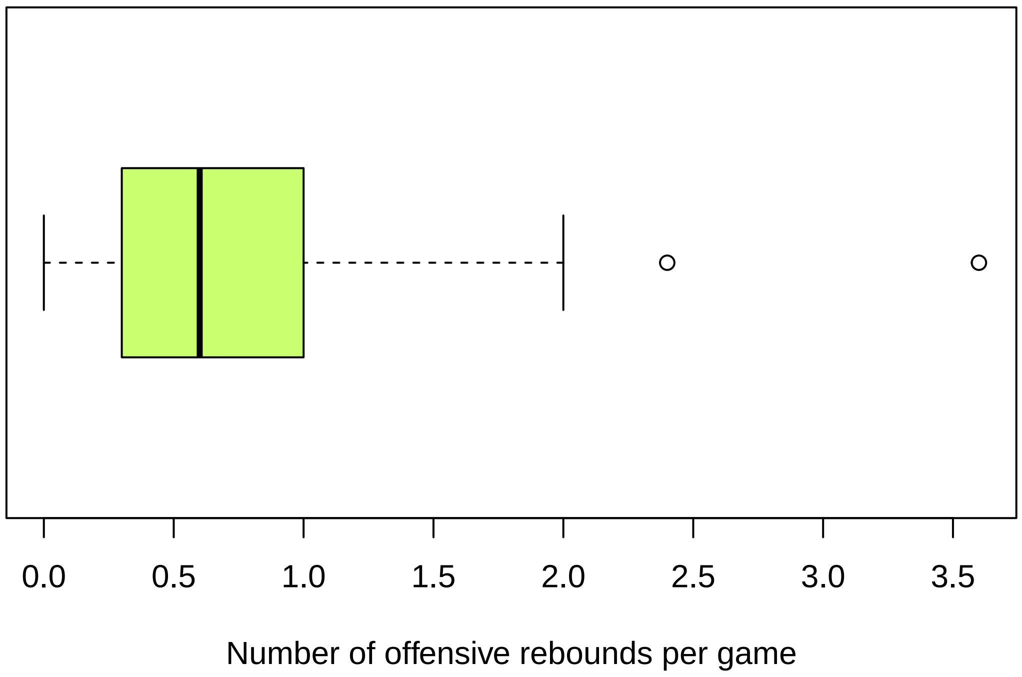
```r
fivenum(NBA$FGA)[2] - 1.5 * IQR # Lower Fence
```

```
## [1] -5.2
```

Part (e) [1 mark]

Make a horizontal outlier boxplot of the `ORB` (average number of offensive rebounds per game) variable. Give an appropriate title to the x-axis and change the color to `darkolivegreen1`.

```r
boxplot(NBA$ORB, horizontal = TRUE, col = "darkolivegreen1",
        xlab = "Number of offensive rebounds per game")
```



Part (f) [1 mark]

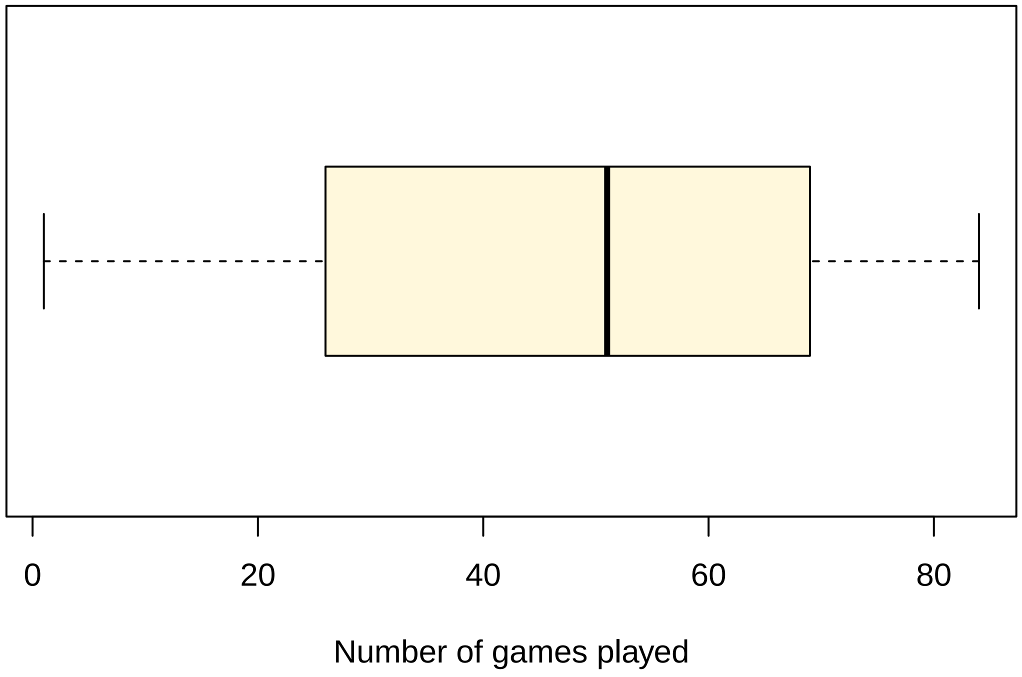
How many players' offensive rebounds per game are considered outliers? (You don't need to do any calculations to answer this.)

2

Part (g) [1 marks]

Make a horizontal outlier boxplot of the G (number of games played) variable. Give an appropriate title to the x-axis and change the color to `cornsilk`.

```
boxplot(NBA$G, xlab = "Number of games played",
        col = "cornsilk", horizontal = TRUE)
```



Part (h) [2 marks]

Do you expect the means of the ORB and G variables to be above or below the respective medians? Explain your answer.

ORB - mean is above the median (right skew). G - mean is below median (left skew).

Part (i) [2 marks]

Confirm your responses above by calculating the means and medians of the ORB and G variables in R.

```
mean(NBA$ORB)
```

```
## [1] 0.6926209
```

```
median(NBA$ORB)
```

```
## [1] 0.6
```

```
mean(NBA$G)
```

```
## [1] 47.78626
```

```
median(NBA$G)
```

```
## [1] 51
```

Part (j) [2 marks]

Would it be more appropriate to summarize the distribution of `ORB` using the mean and standard deviation, or the five-number summary? Explain your answer.

Five number summary is more appropriate to summarize the distribution (the distribution is skewed and contains outliers)

Part (k) [1 mark]

Using `R`, calculate the mean and the variance of the `Age` variable.

```
mean(NBA$Age)
```

```
## [1] 25.62341
```

```
var(NBA$Age)
```

```
## [1] 17.66394
```

Part (l) [1 mark]

In five years, what will be the mean and standard deviation of the ages of this sample of NBA players? You do not need to use `R` for this part; simply give your answer. **Note: Use values rounded to two decimal places.**

mean: 30.62 sd: 4.20

Part (m) [2 marks]

We can use the `table` function to create a frequency distribution of variables. For example, to create a frequency distribution of the `Team` variable, we could use the code `table(NBA$Team)`.

There are five positions in the game of basketball – center (C), power forward (PF), point guard (PG), small forward (SF) and shooting guard (SG). Below, create a frequency distribution of the `Position` variable. Save this table as an object named `positionTable`, and then on the next line, type its name to print it out.

```
positionTable <- table(NBA$Position)
positionTable
```

```
##
##    C   PF   PG   SF   SG
##   48   63   78  100  104
```

For this sample, what is the least frequently observed position?

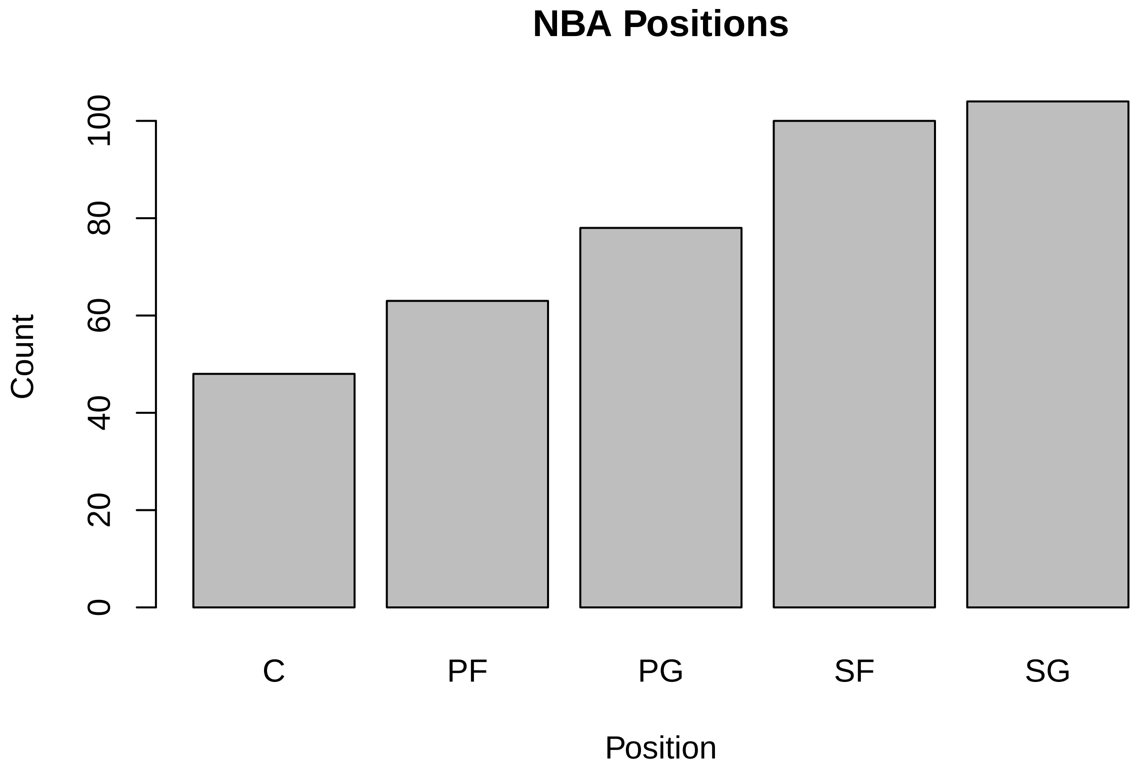
C(center)

Part (n) [2 marks]

Once we have a table, we can use the `barplot` function to create a barplot of a variable. For example, to create a barplot of a table named `myTable`, you would enter `barplot(myTable)`.

Below, create a barplot of the `Position` variable, using the table you created earlier. Set the main title to "NBA Positions", the x-axis label to "Position", and the y-axis label to "Count".

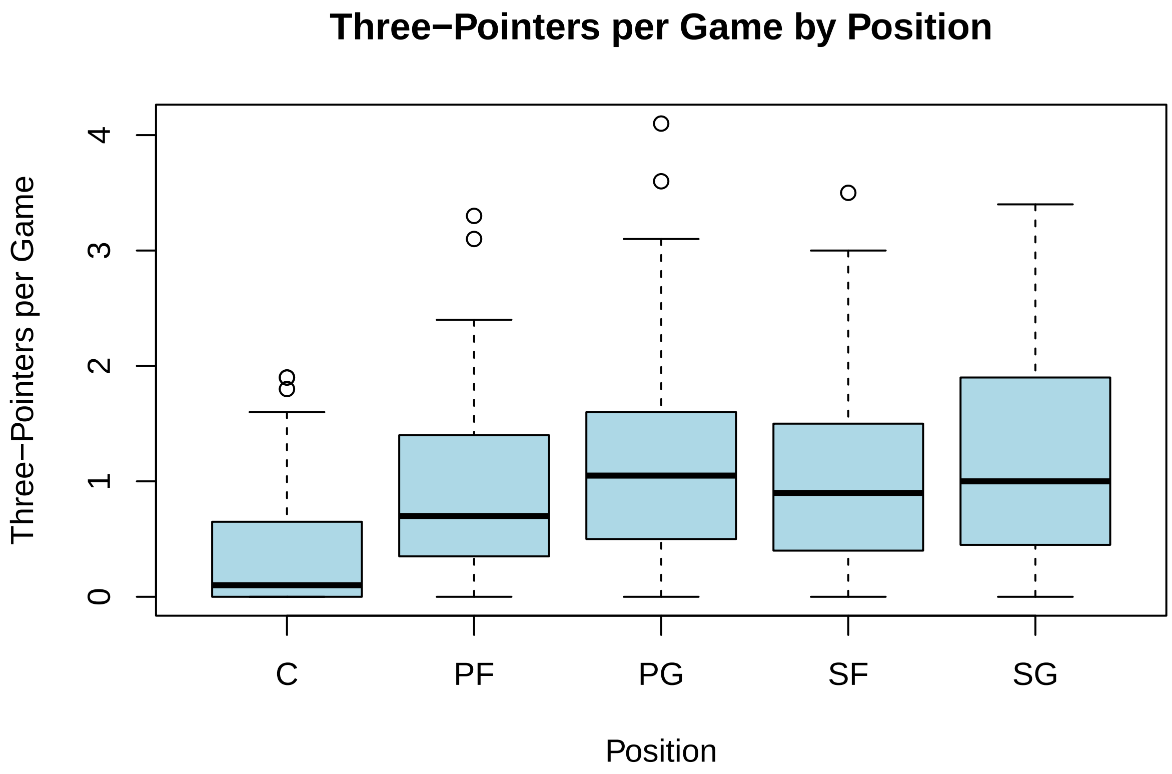
```
barplot(positionTable, main = "NBA Positions", xlab = "Position", ylab = "Count")
```



Part (o) [2 marks]

Create side-by-side quantile boxplots comparing the number of three-pointers per game (`X3P`) for the five positions. Set the x-axis label to "Position" and the y-axis label to "Three-Pointers per Game".

```r
boxplot(X3P ~ Position, data = NBA, xlab = "Position", ylab = "Three-Pointers per Game",
        main = "Three-Pointers per Game by Position", col = "lightblue")
```



Part (p) [3 marks]

Based on the side-by-side boxplots, which position has the highest median for the variable X3P? Which position has the smallest interquartile range? About what percentage of centers (C) have values of X3P below the median for power forwards (PF)?

Highest median - PG Smallest interquartile range - C Values of X3P below the median for PF - about 25%

You will need to knit this file to PDF to properly read the following questions.

Question 2 [4 marks]

Below is a frequency distribution summarizing the final marks of students enrolled in STAT 1150 during the Winter 2025 term.

Mark Range	Frequency
10 – 20	7
20 – 30	8
30 – 40	8
40 – 50	9
50 – 60	16
60 – 70	16
70 – 80	18
80 – 90	20
90 – 100	18

Part (a) [1 mark]

What proportion of values fall in the 30 – 40 interval? Use LaTeX formatting with the `\dfrac{}{}` command. Do not use R for this part. Show your work, and round your answer to three decimal places.

$$\dfrac{8}{120} = 0.067$$

Part (b) [1 mark]

What is the shape of the distribution of ages for this final marks?

Left skew

Part (c) [2 marks]

Which interval contains the third quartile of final marks for this term? You only need to show how you find the position of the third quartile. After that, you can simply report your final answer without showing further work.

Q3 position = (3/4) × (n + 1) = 3/4 x 121 = 90.75

So, 80 – 90 interval

Question 3 [7 marks]

Students at the University of Manitoba often visit nearby cafes during breaks or after classes. On a student review platform called ICAS, students leave ratings for these cafes on a scale from 1 to 5 stars. One such cafe, Sugar Bear, has received the following 270 student ratings:

Rating	Frequency
5	100
4	80
3	50
2	25
1	15

Part (a) [2 marks]

What is the average student rating for Sugar Bear cafe? Use LaTeX formatting with the `\dfrac{}{}` command. Do not use R for this part. Show your work, and round your answer to three decimal places.

$$\dfrac{(5 \times 100) + (4 \times 80) + (3 \times 50) + (2 \times 25) + (1 \times 15)}{270} = \dfrac{500 + 320 + 150 + 50 + 15}{270} = \dfrac{1035}{270} = 3.833$$

Part (b) [2 marks]

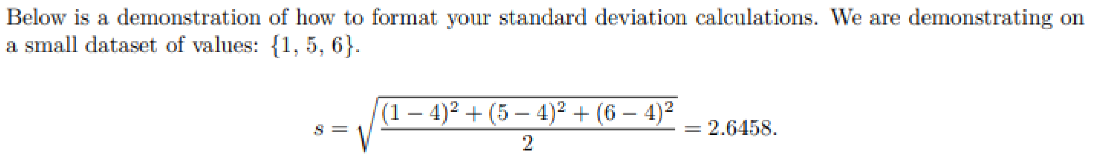
Before Sugar Bear became popular, many students used to go to another nearby spot, IQ Cafe, which had 160 student reviews with an average rating of 3.95. What is the combined average rating for Sugar Bear and IQ Cafe? Use LaTeX formatting with the `\dfrac{}{}` command. Do not use R for this part. Show your work, and round your answer to three decimal places.

$$\text{Mean} = \dfrac{(270 \times 3.833) + (160 \times 3.95)}{270 + 160} = \dfrac{1034.91 + 632}{430} = \dfrac{1666.91}{430} = 3.876$$

Part (c) [2 marks]

Four students left new ratings for Sugar Bear during the past week, giving the following scores: 4, 5, 3, and 4. What is the sample standard deviation of these four new ratings? When formatting your answer in LaTeX, follow the same formatting as the following example. (Knit this document now to see the equation.)

Suppose we have a small data set consisting of the values 1, 5 and 6. The standard deviation for this data set is



Below is a demonstration of how to format your standard deviation calculations. We are demonstrating on a small dataset of values: $\{1, 5, 6\}$.

$$s = \sqrt{\dfrac{(1-4)^2 + (5-4)^2 + (6-4)^2}{2}} = 2.6458.$$

Calculate the standard deviation for Sugar Bear's four new ratings. Do not show any more, or fewer, steps than are presented in the equation above. Round your answer to four decimal places.

$$s = \sqrt{\frac{(4-4)^2 + (5-4)^2 + (3-4)^2 + (4-4)^2}{4-1}} = 0.8165$$

Part (d) [1 marks]

Create a vector of these four data values called `Rating`, and then use R to verify your calculation in part (c).

```r
Rating <- c(4, 5, 3, 4)
sd(Rating)
```

```
## [1] 0.8164966
```

Question 4 [3 marks]

Part (a) [1 mark]

You roll a 6-sided die (numbered 1 through 6) and a 4-sided die (numbered 1 through 4). The outcome of interest is the ordered pair showing the result of each die roll (first the 6-sided die, then the 4-sided die). Write out the complete sample space of outcomes for this experiment.

{ (1,1), (1,2), (1,3), (1,4), (2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4), (4,1), (4,2), (4,3), (4,4), (5,1), (5,2), (5,3), (5,4), (6,1), (6,2), (6,3), (6,4) }

Part (b) [1 mark]

Suppose we are interested in the event A that the 6-sided die shows a higher number than the 4-sided die. List all the outcomes in the event A

{ (2,1), (3,1), (3,2), (4,1), (4,2), (4,3), (5,1), (5,2), (5,3), (5,4), (6,1), (6,2), (6,3), (6,4) }

Part (C) [1 mark]

What does the complement event A^c represent?

{(1,1),(1,2),(1,3),(1,4),(2,2),(2,3),(2,4),(3,3),(3,4),(4,4)}

Question 5 [4 marks]

You will need to knit the file in order to properly view this question.

Part (a) [2 marks]

There are eight countries competing in an international men's cricket World Cup. The probability of each team winning the tournament is shown in the table below, where k is some constant:

Team	Australia	Sri Lanka	India	Bangladesh	England	New Zealand	South Africa	West Indies
Prob.	$2k$	0.09	$2k$	0.11	k	0.15	0.09	0.06

What is the probability that an Asian country wins the tournament? (The Asian teams are Sri Lanka, India, and Bangladesh.) Show your work. (Type your answer below. In order for your answer to appear in italics, you need to type your text between asterisks.)

P(Asian country wins) = 0.09 + 2k + 0.11 = 0.09 + 0.20 + 0.11 = 0.40

Part (b) [2 marks]

There are also eight countries competing in an international women's cricket World Cup. The probability of each team winning the tournament is shown in the table below, with some probabilities missing:

Team	Australia	Sri Lanka	India	Bangladesh	England	New Zealand	Pakistan	West Indies
Prob.	0.25	???	???	0.20	0.10	0.08	0.10	0.05

What is the probability that an Asian country wins the tournament? (The Asian teams are Pakistan, India, Sri Lanka, and Bangladesh.) Show your work. (Type your answer below. In order for your answer to appear in italics, you need to type your text between asterisks.)

P(Asian country wins) = 0.10 + 0.20 + x + y = 0.30 + 0.22 = 0.52

Question 6 [14 marks]

The GRE is a standardized test used for admission into many PhD programs worldwide. The quantitative section of the GRE is scored out of 170 points.

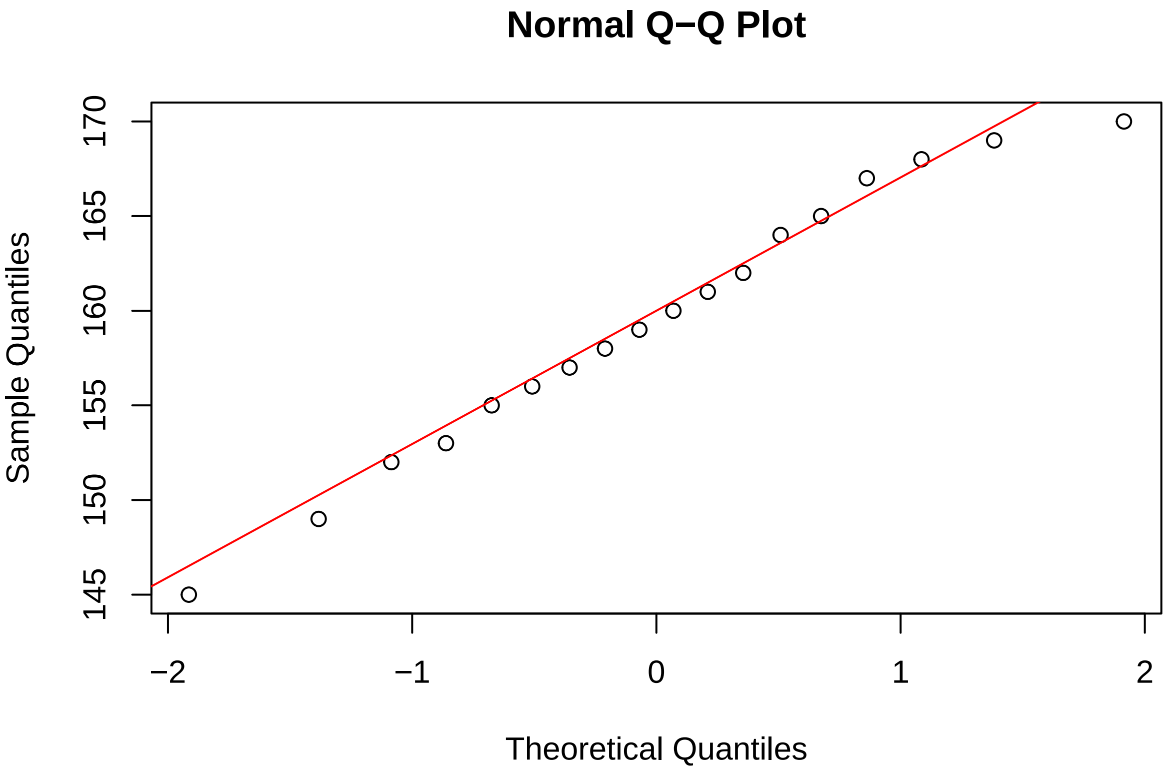
Part (a) [1 mark]

The GRE Quantitative Reasoning scores of a random sample of 18 students applying to PhD programs are shown below. Run the following line of code to import this data set as a vector named `GRE`:

```
GRE <-  c(145, 149, 152, 153, 155, 156, 157, 158, 159, 160,
              161, 162, 164, 165, 167, 168, 169, 170)
```

Create a normal quantile plot for this data set, and add a reference line in red.

```
qqnorm(GRE)
qqline(GRE, col = "red")
```



Part (b) [1 mark]

Does it appear reasonable to assume that SAT scores follow a normal distribution? Why or why not?

The points in the normal quantile plot follow a straight line, so - SAT scores follow a normal distribution

For the remainder of this question, assume that GRE Quantitative scores follows a normally distributed with mean of 155 and standard deviation of 6.

Part (c) [2 marks]

What is the probability that a randomly selected student scores less than 165 on the GRE Quantitative section? Use LaTeX formatting and refer to the normal table to find your answer. Show your work. (Do not use R for this part.) Your work should look similar to the solution for Question 11(a) in the Unit 2 Practice Problems on UM Learn.

$$z = \frac{165 - 155}{6} = \frac{10}{6} \approx 1.667$$

Part (d) [1 mark]

Use R to verify your answer for Part (c). (Note that the probability given by R will be slightly different than the one you calculated above, as R carries more decimal places.)

```r
pnorm(165, mean = 155, sd = 6)
```

```
## [1] 0.9522096
```

Part (e) [2 marks]

What is the probability that a randomly selected student has a GRE Quantitative scores greater than 150? Use LaTeX formatting and refer to the normal table to find your answer. Show your work. (Do not use R for this part.) Your work should look similar to the solution for Question 11(b) in the Unit 2 Practice Problems on UM Learn.

$$z = \frac{150 - 155}{6} = \frac{-5}{6} \approx -0.833$$
$$P(X > 150) = P(Z > -0.83) = 1 - P(Z < -0.83) = 1 - 0.2033 = 0.7967$$

Part (f) [1 mark]

Use R to verify your answer for Part (e). (Note that the probability given by R will be slightly different than the one you calculated above, as R carries more decimal places.)

```r
pnorm(150, mean = 155, sd = 6, lower.tail = FALSE)
```

```
## [1] 0.7976716
```

Part (g) [1 mark]

Use R to find the probability that a randomly selected student has a GRE Quantitative scores between 140 and 149.

```r
pnorm(149, mean = 155, sd = 6) - pnorm(140, mean = 155, sd = 6)
```

```
## [1] 0.1524456
```

Part (h) [1 mark]

The top 13% of students have GRE Quantitative scores above what value? Use R to find your answer.

```
qnorm(1 - 0.13, mean = 155, sd = 6)
```

```
## [1] 161.7583
```

Part (i) [2 marks]

What is the interquartile range of GRE Quantitative scores? Use R to find your answer.

```
IQR <- fivenum(GRE)[4] - fivenum(GRE)[2]
IQR
```

```
## [1] 10
```

Part (j) [2 marks]

One student scored 162 on the GRE Quantitative section and 5.0 on the GRE Analytical Writing section, which follows a normal distribution with mean 3.9 and standard deviation 0.7. On which section did the student perform better relative to peers? Use LaTeX formatting and show your work. (Type your answer below. Your LaTeX work will be between double dollar signs. In order for the text in your answer to appear in italics, you need to type your text between asterisks.)

Calculate z-scores: **GRE Quantitative:**

$$z = \frac{162 - 155}{6} = \frac{7}{6} = 1.167$$

GRE Analytical Writing:

$$z = \frac{5.0 - 3.9}{0.7} = \frac{1.1}{0.7} = 1.571$$

z-score is higher for Analytical Writing (1.571) than for Quantitative (1.167), so the student performed better Analytical Writing section.

Question 7 [5 marks]

`R` can also be used to calculate probabilities for a variable that follows a uniform distribution.

Suppose a variable X follows a uniform distribution on the interval from `a` to `b`.

To find $P(X < x)$, we use the command `punif(x, a, b)`.

To find $P(X > x)$, we can either use the command
`punif(x, a, b, lower.tail = FALSE)` or `1 - punif(x, a, b)`.

To find $P(x1 < X < x2)$, we use the command `punif(x2, a, b) - punif(x1, a, b)`.

To find the p^{th} percentile of the distribution of X (i.e., the value x such that $P(X < x) = p$), we use the command `qunif(p, a, b)`.

The time it takes for an electronic scooter to fully charge follows a uniform distribution between 1.5 and 6.5 hours. Use R for each part of this question, except Part (d).

Part (a) [1 mark]

What is the probability it takes less than 120 minutes to fully charge the scooter?

```
punif(2, 1.5, 6.5)
```

```
## [1] 0.1
```

Part (b) [1 mark]

What is the probability it takes more than 3.1 hours to fully charge the scooter?

```
punif(3.1, 1.5, 6.5, lower.tail = FALSE)
```

```
## [1] 0.68
```

Part (c) [1 mark]

What is the probability it takes between 3.7 and 5.9 hours to fully charge the scooter?

```
punif(5.9, 1.5, 6.5) - punif(3.7, 1.5, 6.5)
```

```
## [1] 0.44
```

Part (d) [1 mark]

What is the probability it takes exactly 4 hours to charge the scooter?

0

Part (e) [1 mark]

In 12.5% of charging sessions, it takes less than how many hours to fully charge the scooter?”

```r
qunif(0.125, 1.5, 6.5)
```

```
## [1] 2.125
```