Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение

высшего образования

**«КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»**

**(ФГБОУ ВО «КубГУ»)**

**Факультет компьютерных технологий и прикладной математики**

**Кафедра вычислительных технологий**

**Отчёт лабораторная №6**

**Дисциплина: Обработка больших данных**

Работу выполнил \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Логвина А.В.

Направление подготовки 02.03.02 Фундаментальная информатика и

информационные технологии

Преподаватель \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Яхонтов А.А.

Краснодар

2025

**Проверка статистических гипотез.**

**Цель работы:** ознакомиться с некоторыми статистическими тестами, принципами их работы. Научиться оценивать нормальность распределения выборки, а также выполнять оценку статистических гипотез.

**Задание:**

1. Провести дескриптивный анализ данных.

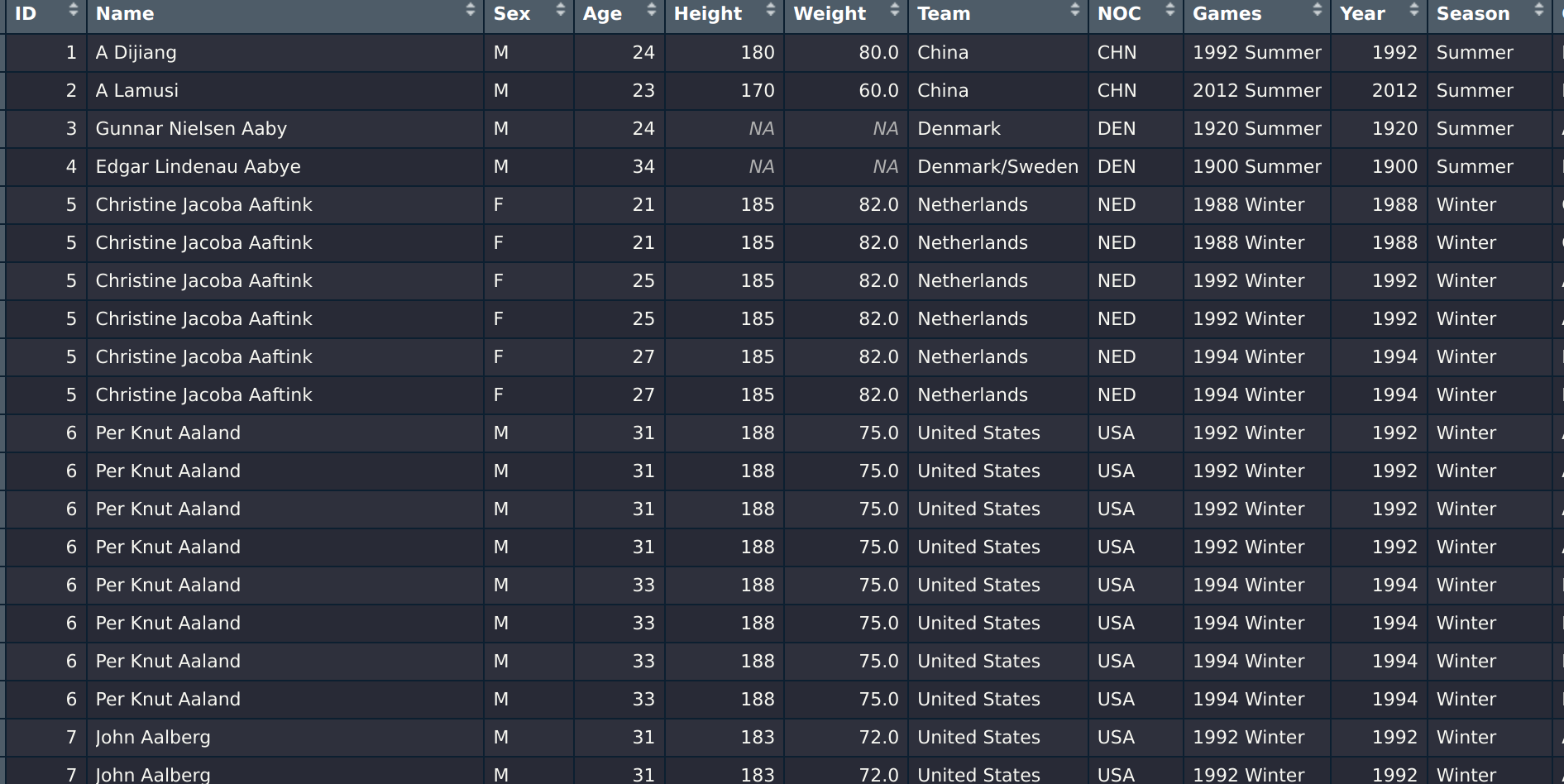
2. Провести проверку на нормальность и дисперсию. Из чего сделать вывод о требуемом тесте.

3. Проверьте гипотезу о среднем весе спортсменов выбранного вида спорта (вид спорта остается по ЛР4).

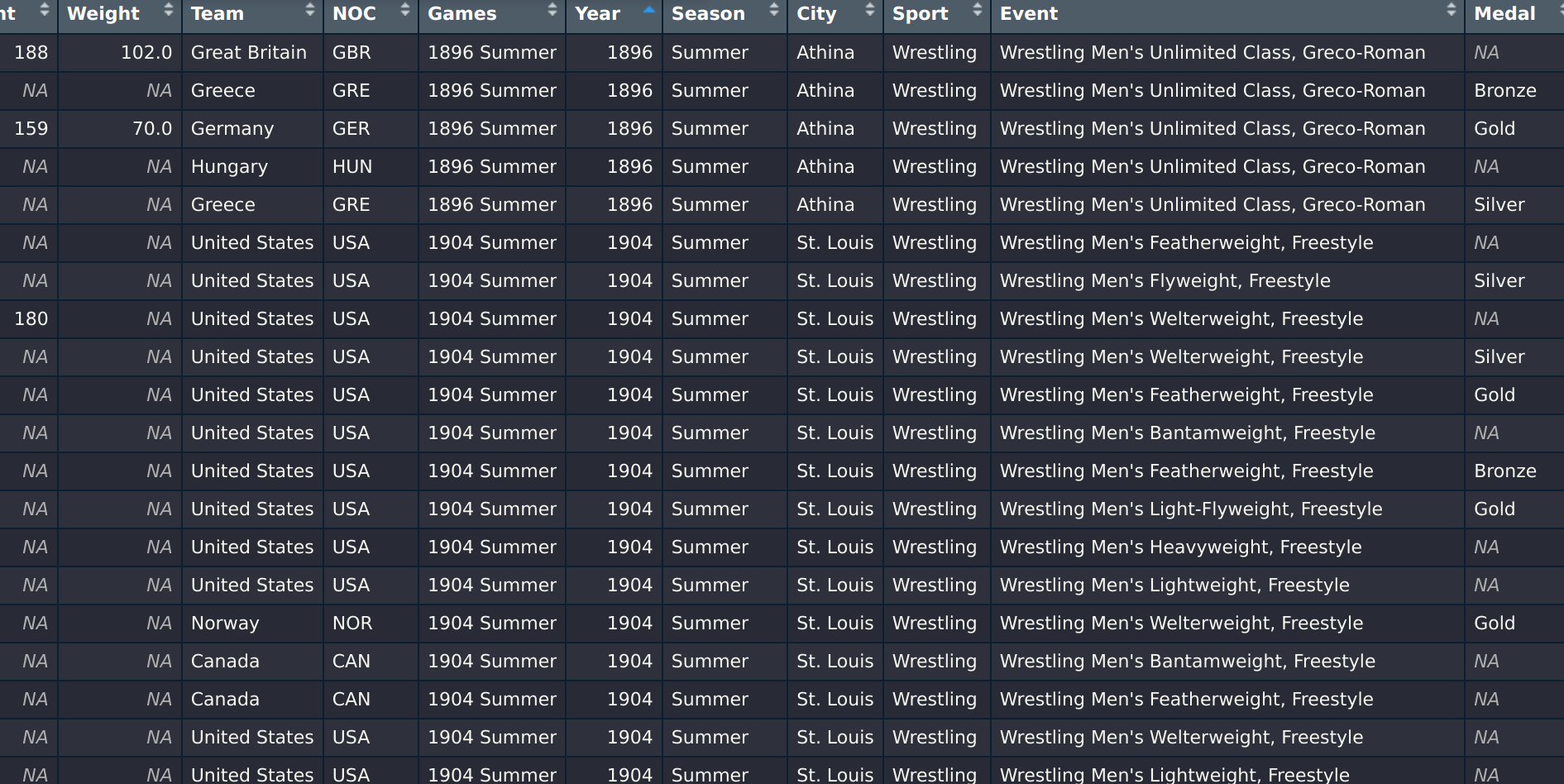
4. Проверьте гипотезу о равенстве среднего веса женщин (мужчин) в двух разных выбранных видах спорта (сравнение двух независимых выборок – двухвыборочный критерий.).

**Ход работы:**

Скачаем датасет:



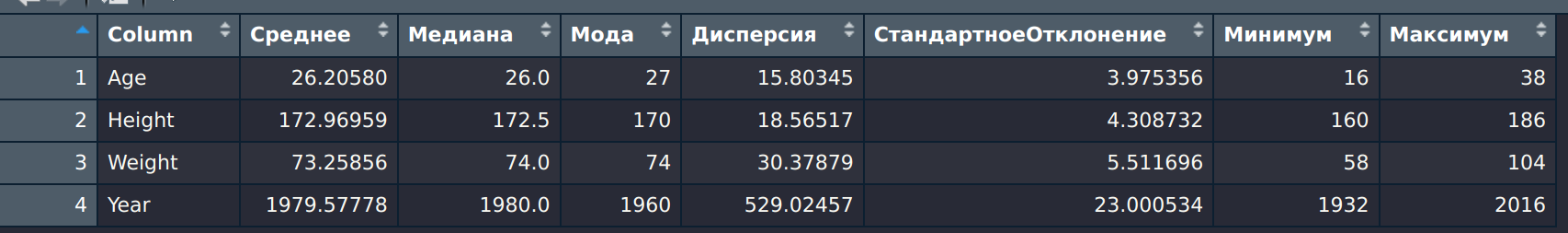
Сделаем датасет только для моего варианта:



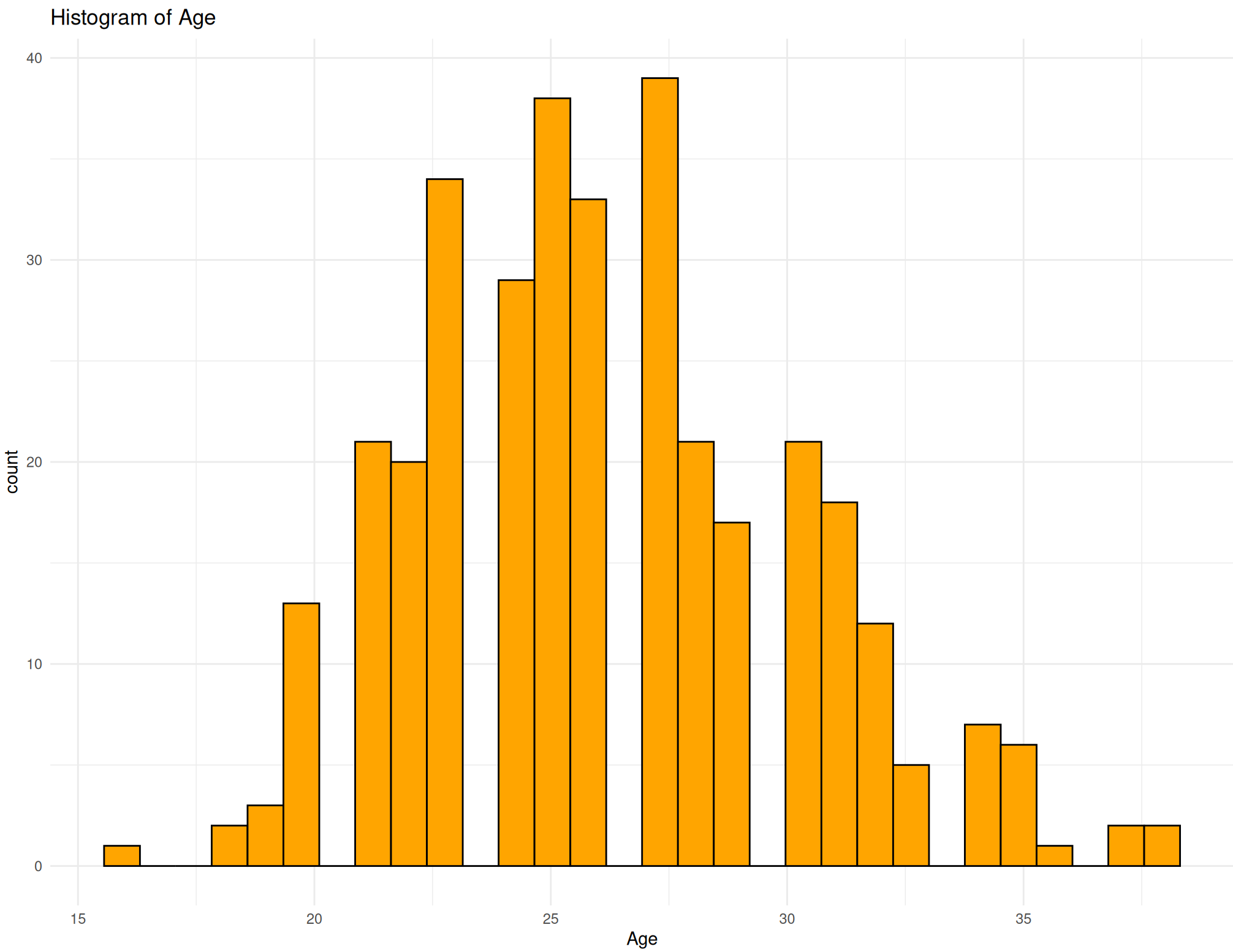
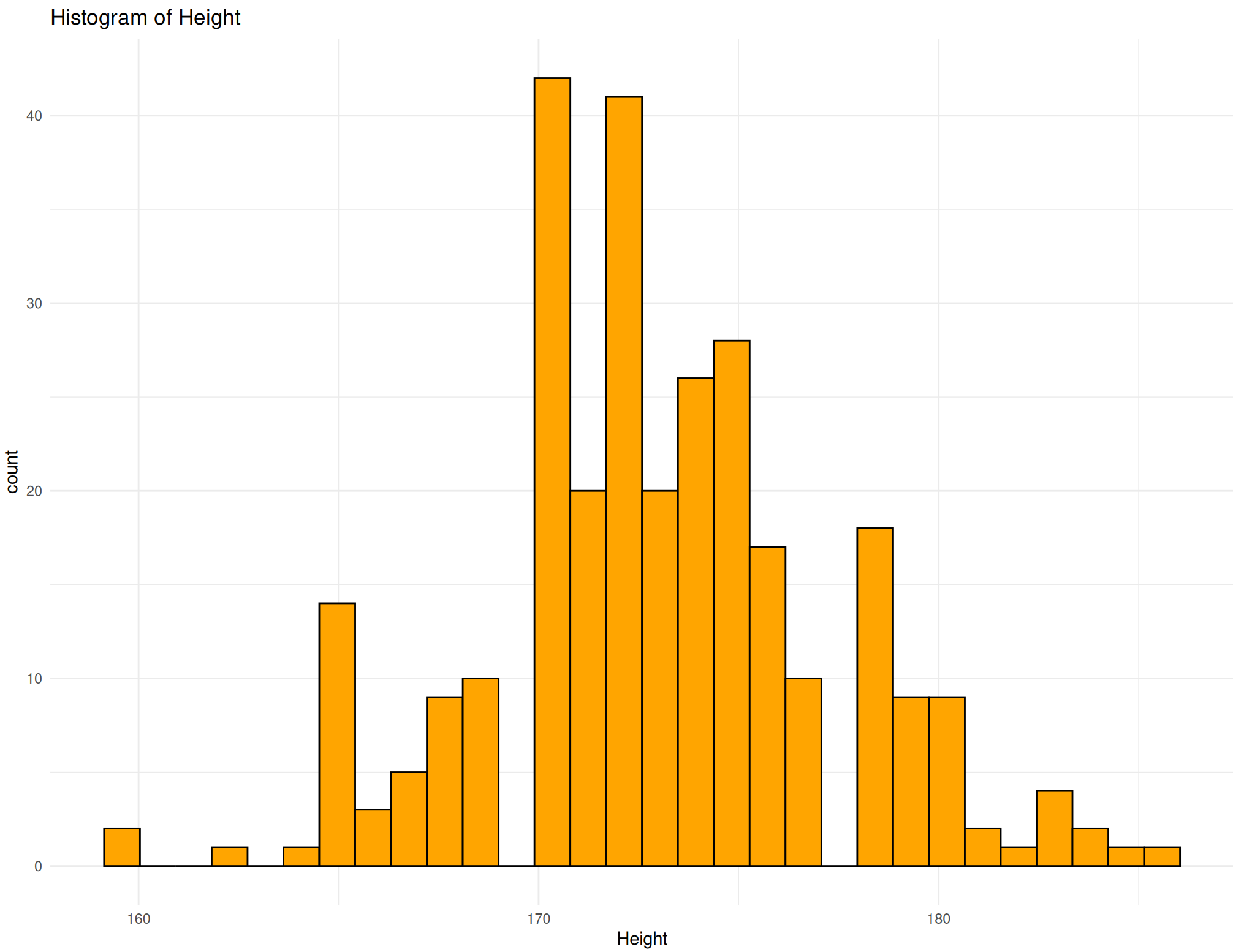
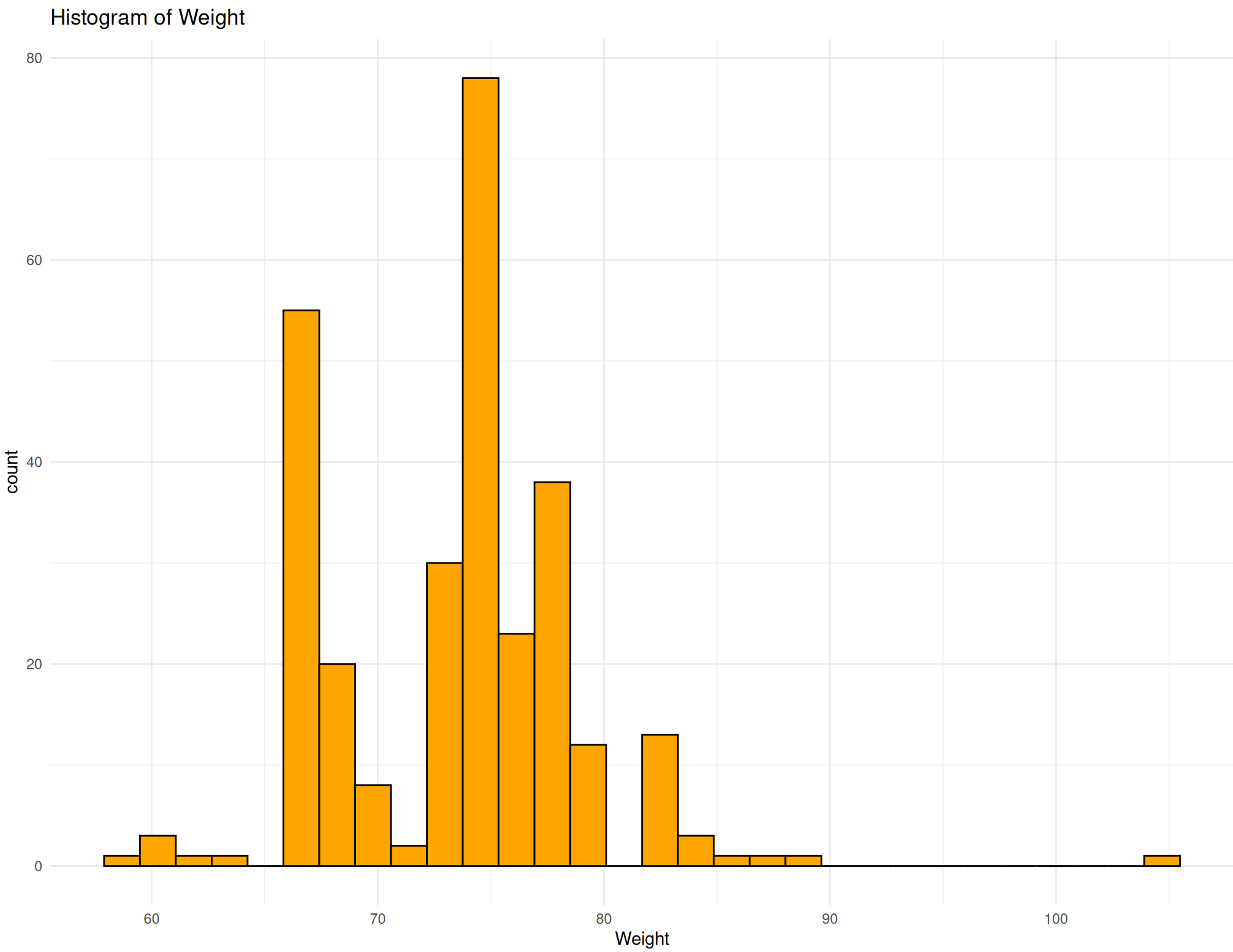
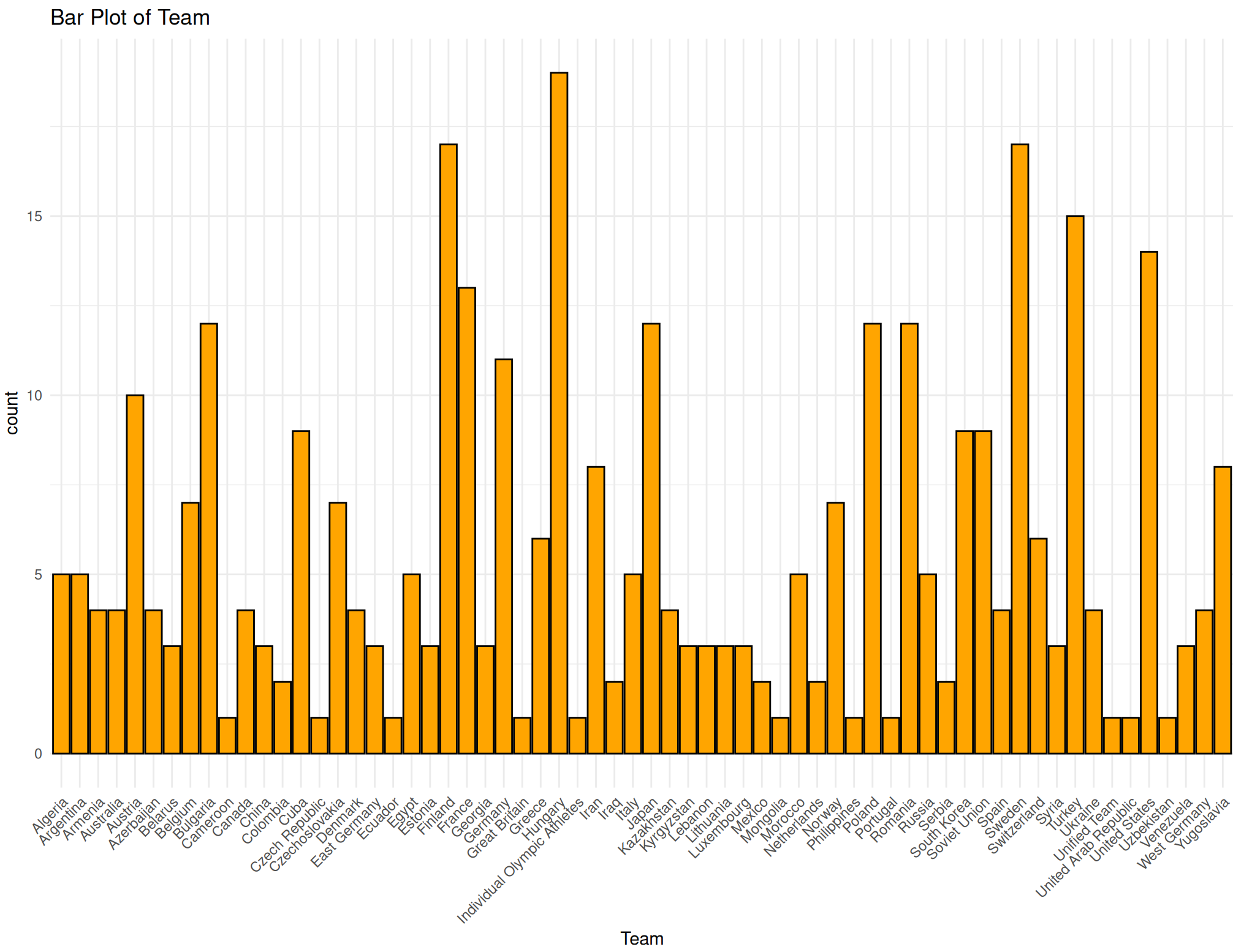
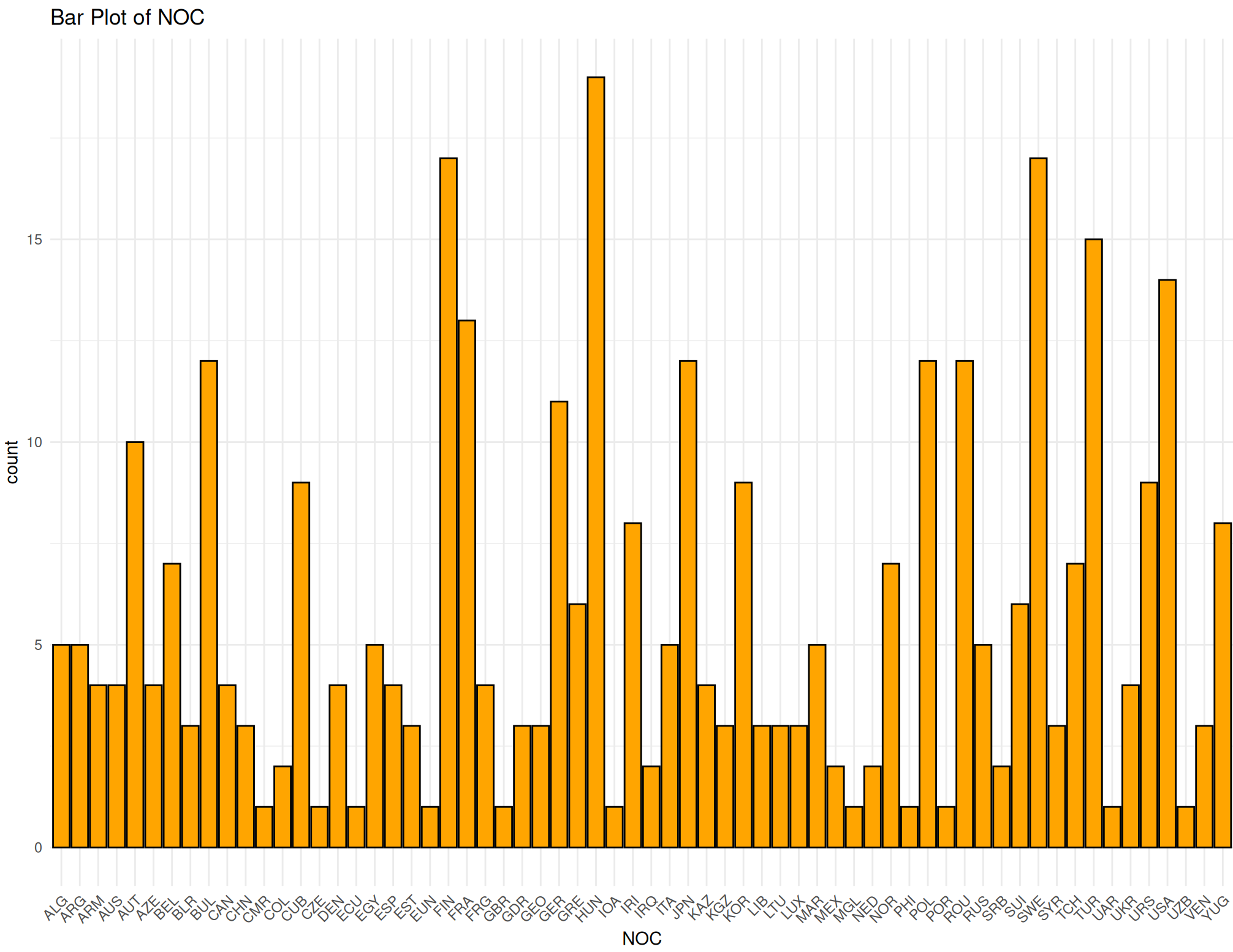
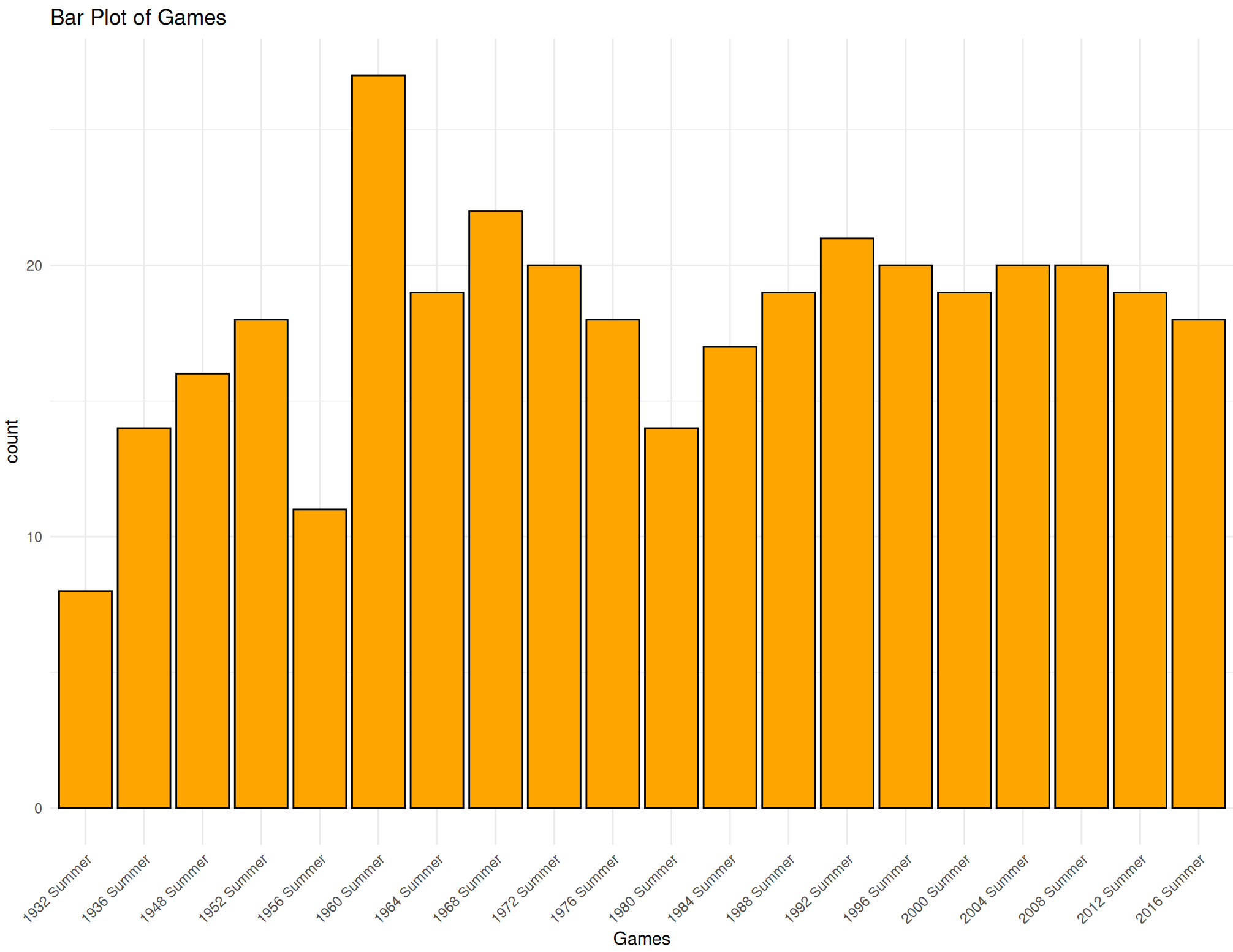
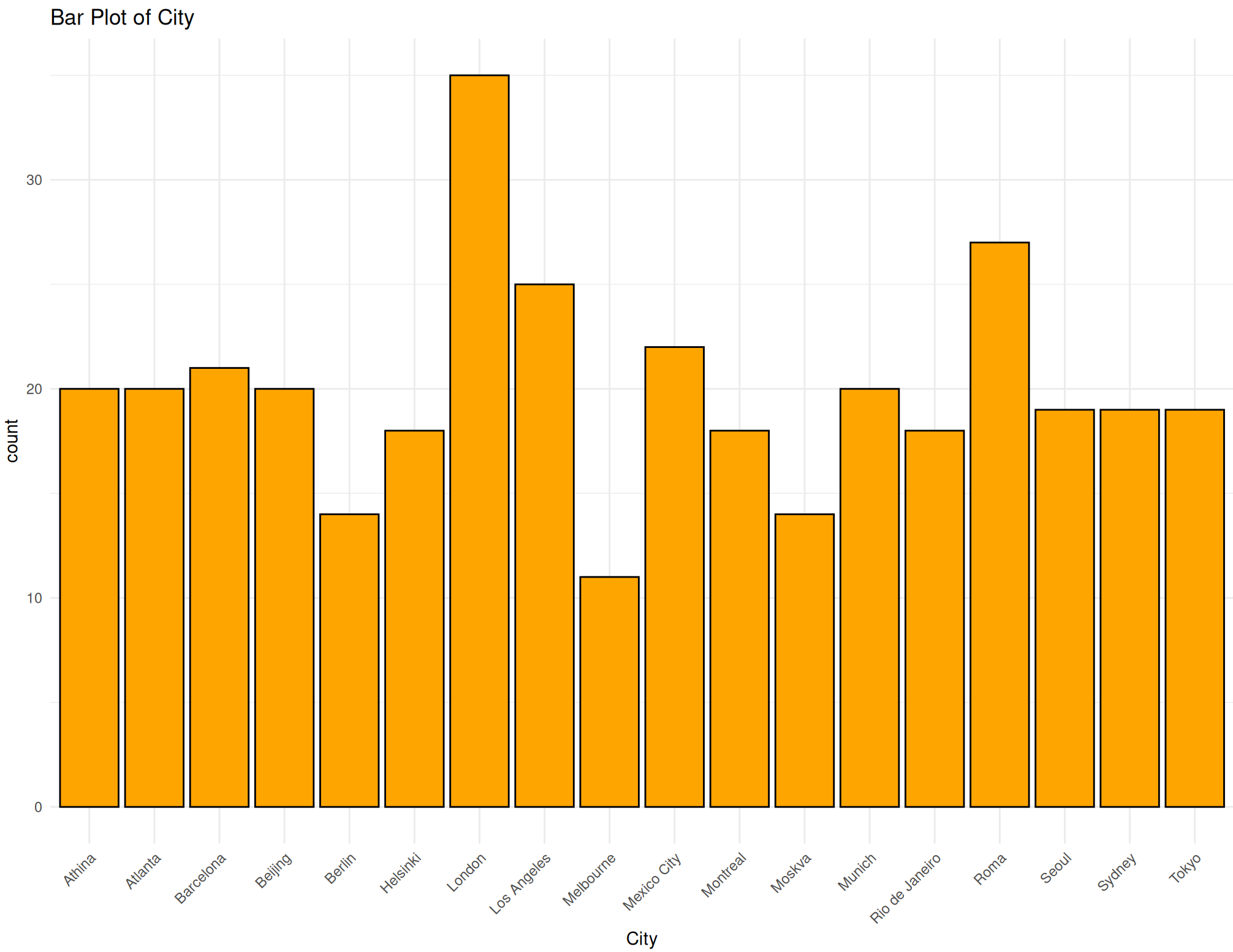
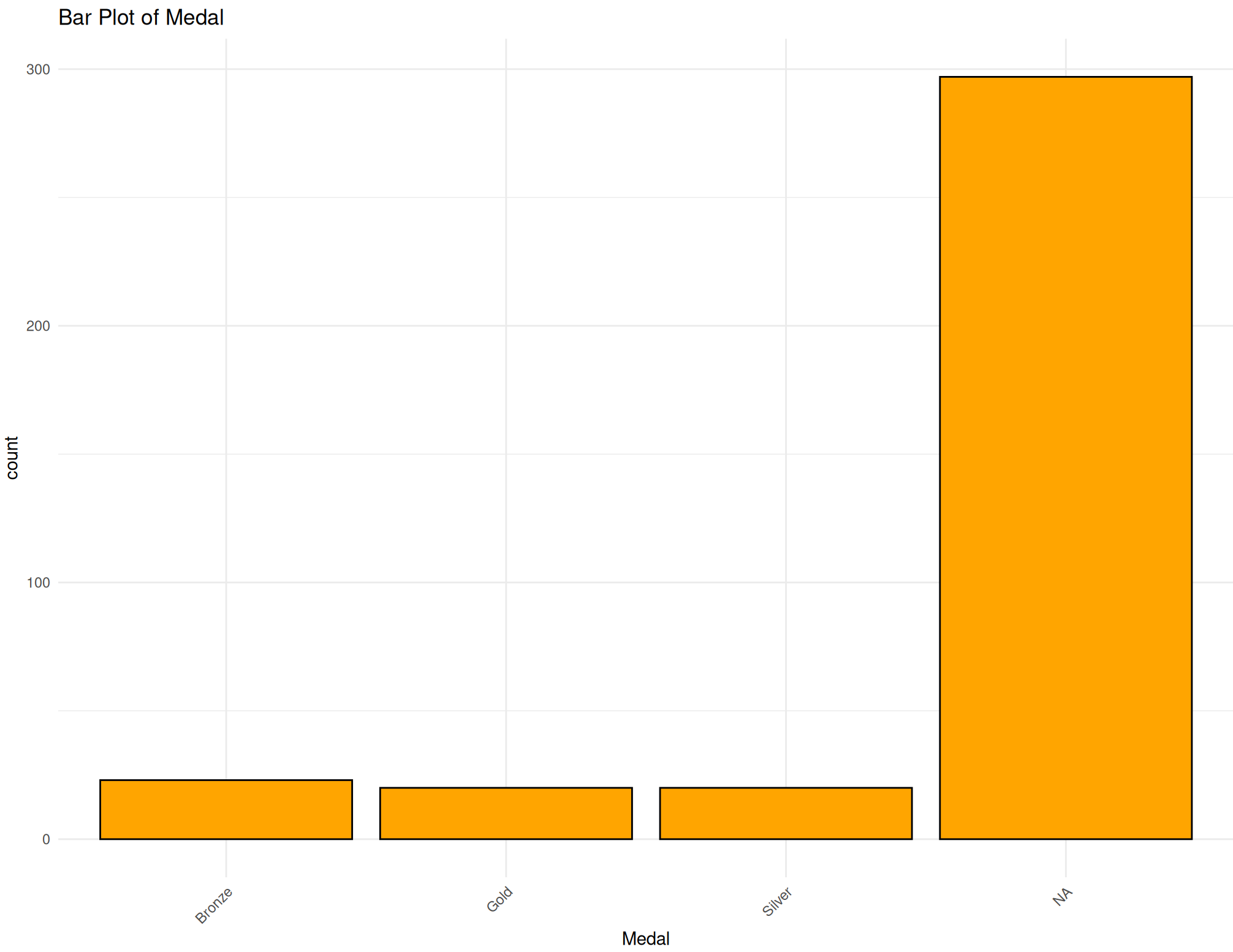
1. Провести дескриптивный анализ данных.

Для этого находим: среднее значение, медиану, моду,дисперсию, стандартное отклонение, минимум, максимум.

Результат анализа:

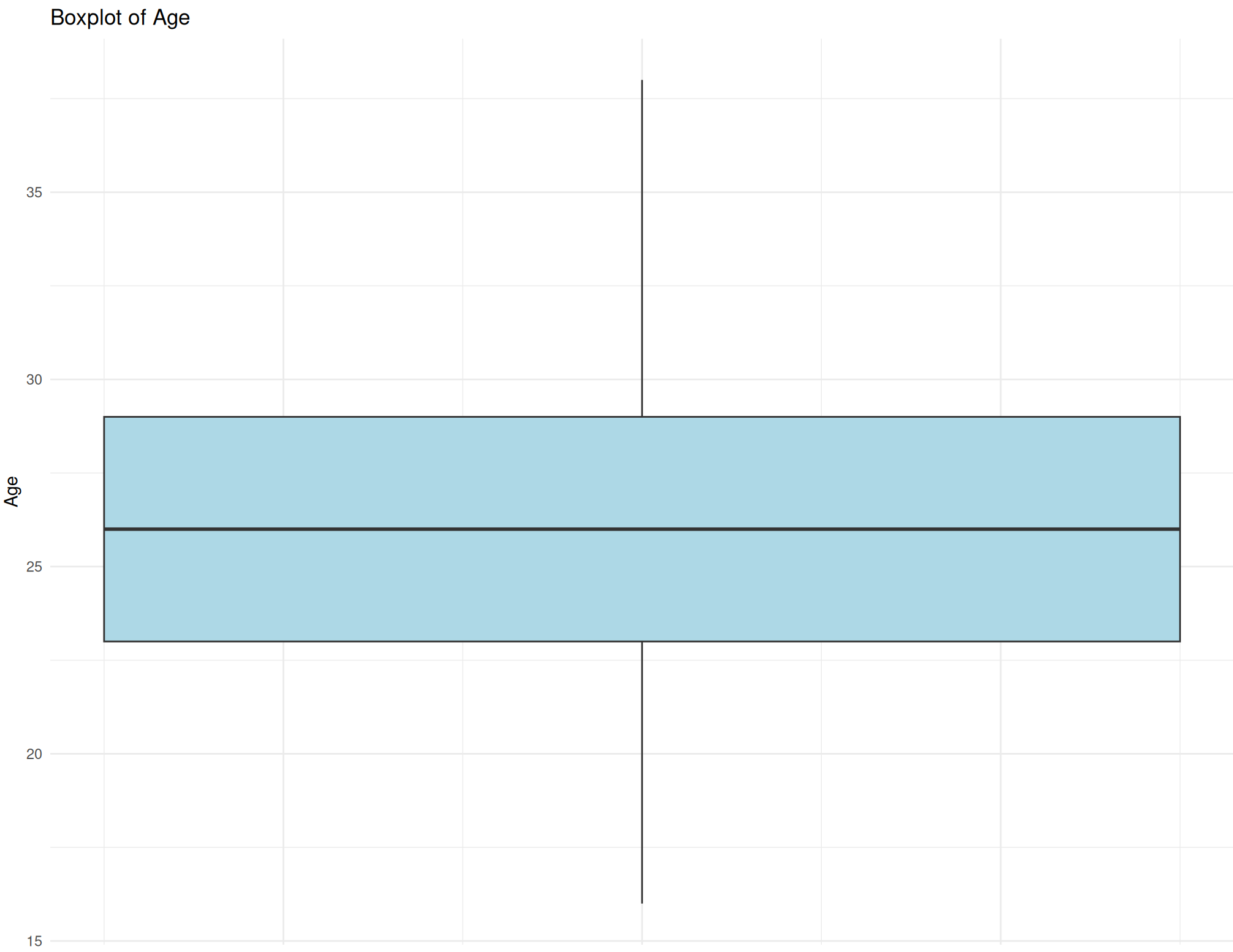
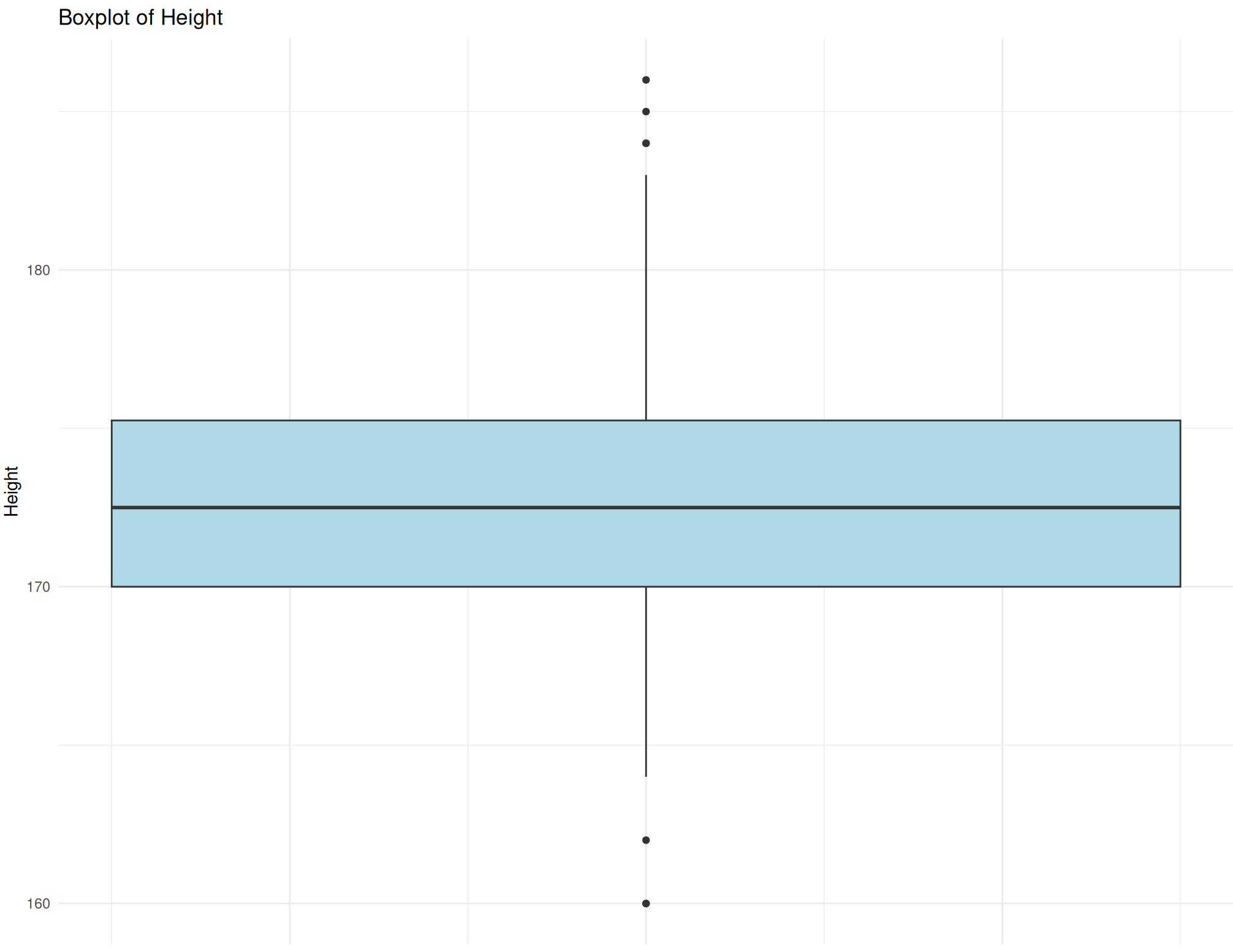
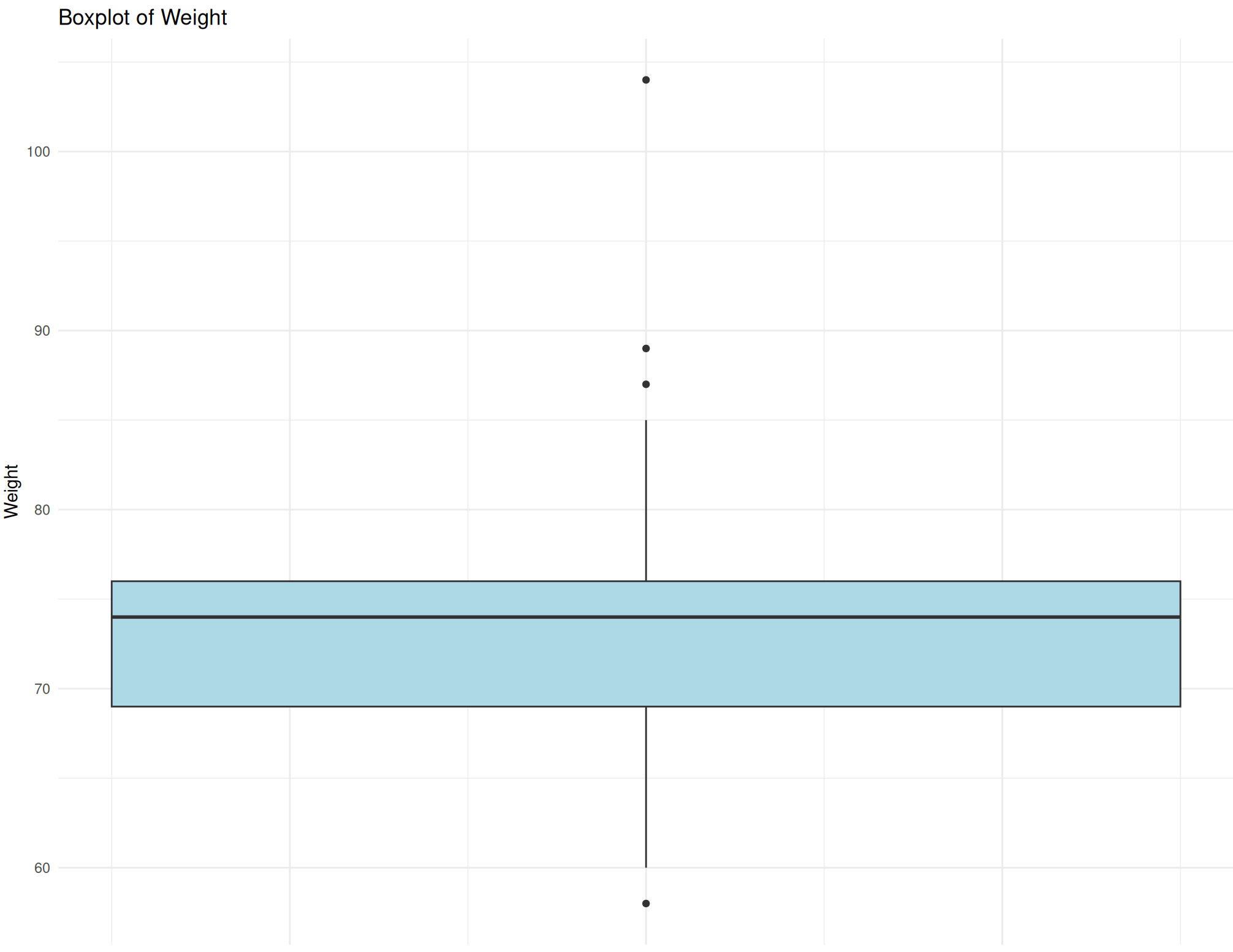


Гистограммы:



Боксплоты:

Боксплоты нельзя строить для нечисловых данных, потому что они требуют вычисления статистик (медианы, квартилей, межквартильного размаха), которые математически неопределимы для текста, так как не имеют числового порядка, а значит, невозможно определить их распределение, разброс или выбросы — ключевые элементы, которые визуализирует боксплот. Попытка применить этот метод к нечисловым данным приведёт к ошибке в программе или бессмысленному графику, так как алгоритм не сможет ранжировать и обрабатывать качественные признаки как количественные.



2. Провести проверку на нормальность и дисперсию. Из чего сделать вывод о требуемом тесте.

Нормальность:

Q-Q график для Year показывает значительные отклонения от красной линии, особенно в средней и нижней частях. Точки формируют выраженный ступенчатый узор, что связано с тем, что годы проведения Олимпийских игр дискретны (например, 1920, 1924 и т.д.), и распределение не может быть нормальным. Это подтверждает, что переменная Year не следует нормальному распределению, что ожидаемо для такого

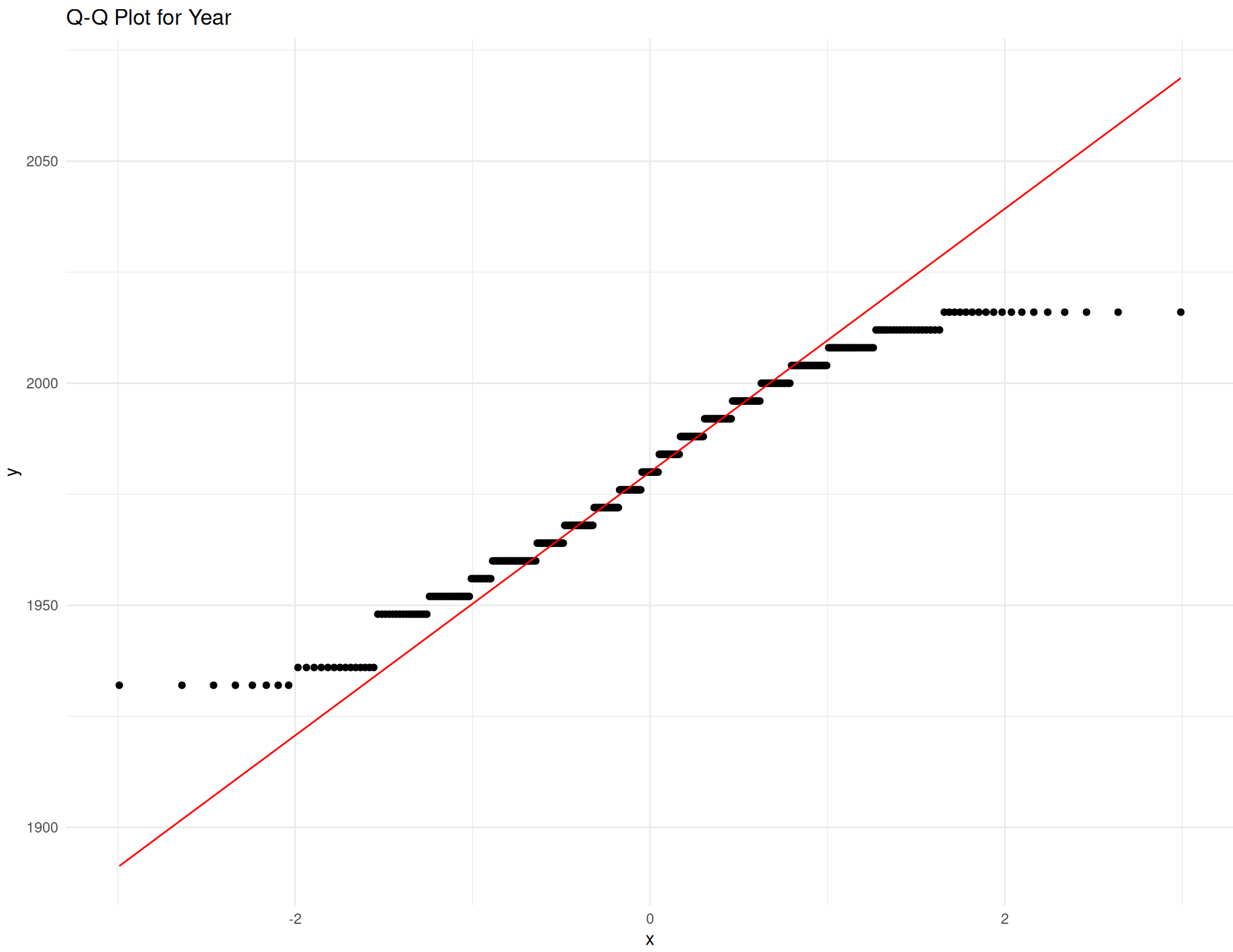
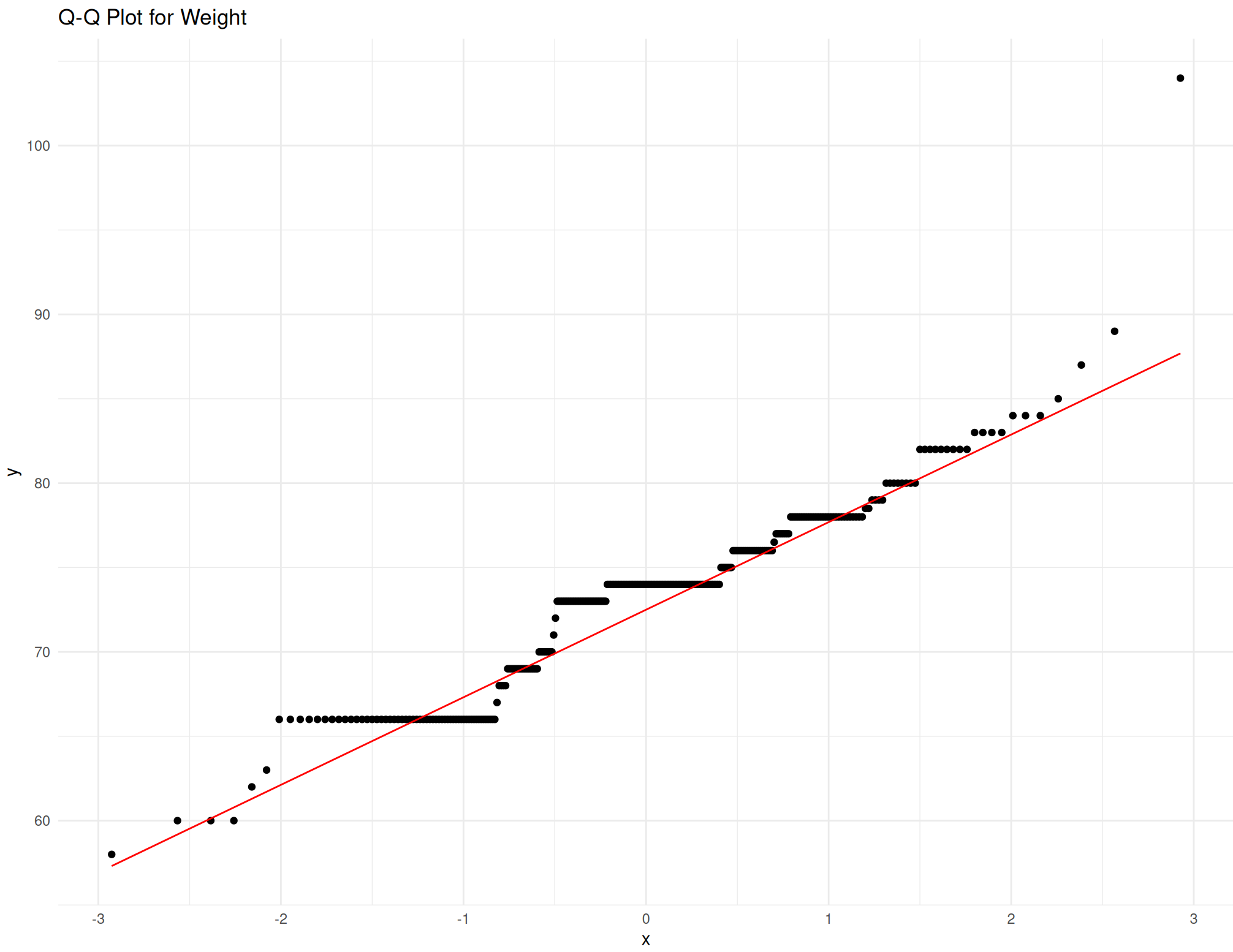
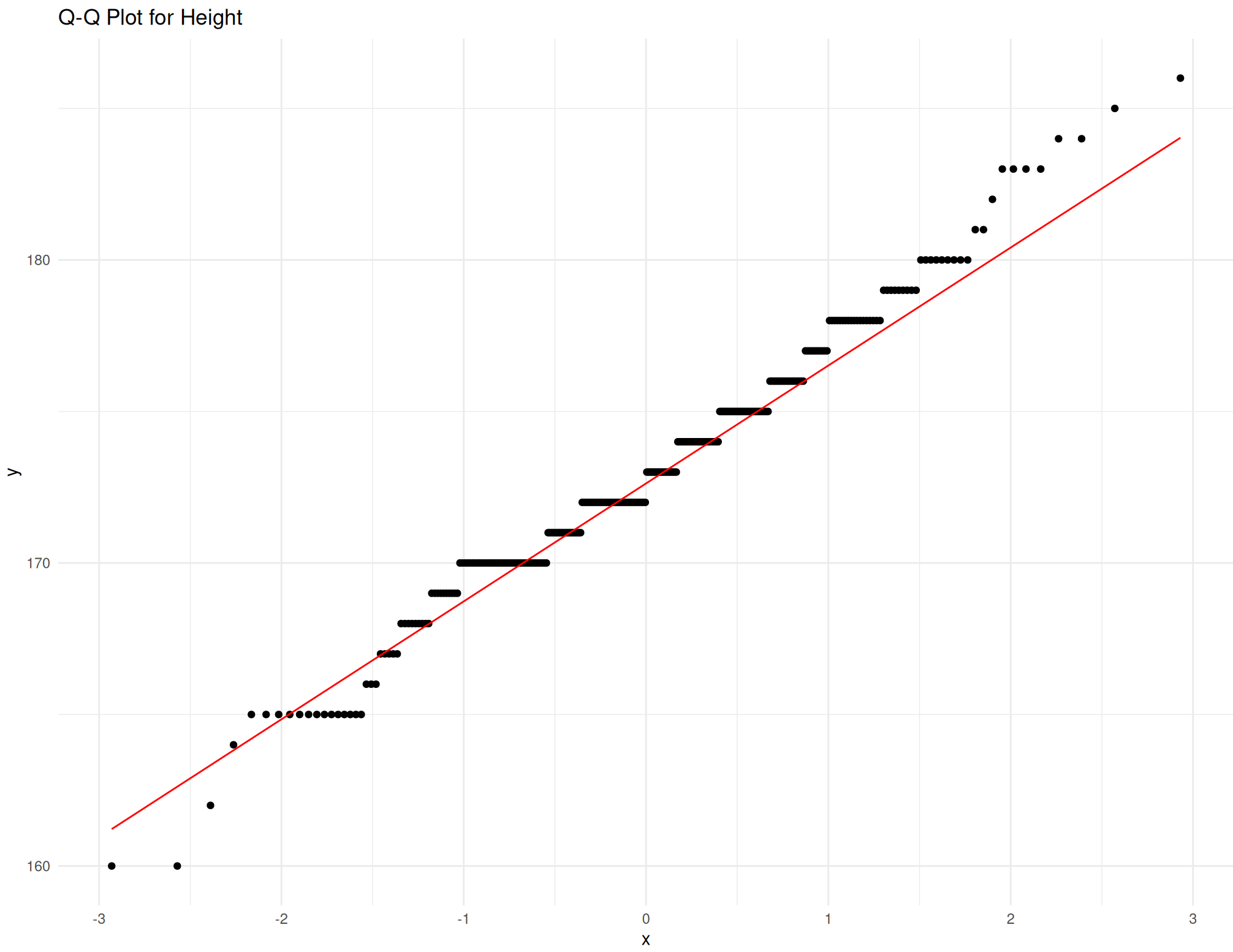


График Q-Q для переменной Weight показывает значительные отклонения точек от красной линии, особенно в нижней и верхней частях распределения. Точки формируют ступенчатый узор, что указывает на дискретность данных и возможное наличие повторяющихся значений, что нехарактерно для нормального распределения. Это предполагает, что вес борцов в данной категории не следует нормальному распределению, что может быть связано с ограничениями весовой категории в соревнованиях.



Q-Q график для Height демонстрирует умеренное соответствие красной линии, но с заметными отклонениями в нижней и верхней частях. Точки также образуют ступенчатый узор, указывая на дискретность данных, что может быть связано с ограниченным диапазоном роста участников в данной весовой категории. Это говорит о том, что распределение роста, скорее всего, не является нормальным, хотя отклонения менее выражены, чем для веса.



На Q-Q графике для Age точки в целом следуют красной линии, но с небольшими отклонениями в нижней части и более заметными в верхней. Ступенчатый узор присутствует, что указывает на дискретность возраста (целые числа) и возможное ограничение возрастного диапазона участников. Это предполагает, что распределение возраста может быть близким к нормальному, но с некоторыми отклонениями, которые стоит учитывать при выборе теста.

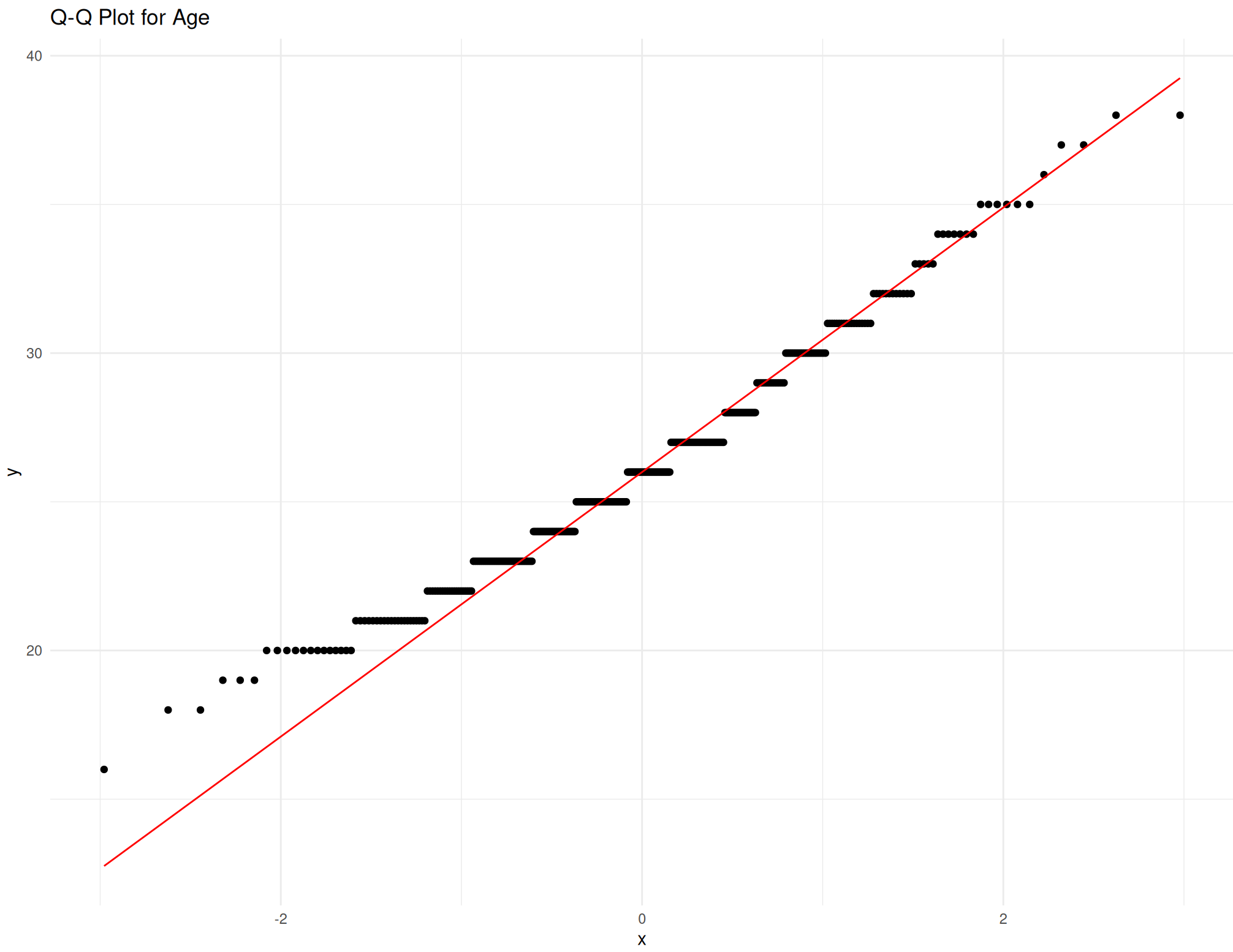


Таблица результатов теста Шапиро-Уилка показывает, что p-значения для всех переменных (Age: 7.66e-05, Height: 3.29e-03, Weight: 2.93e-10, Year: 3.54e-08) значительно меньше 0.05, что указывает на отклонение от нормальности для каждой из них. Статистика W (от 0.932 для Weight до 0.985 для Height) также подтверждает отклонения, так как значения далеки от 1, что было бы характерно для нормального распределения.

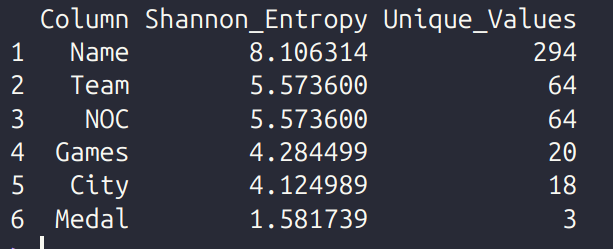
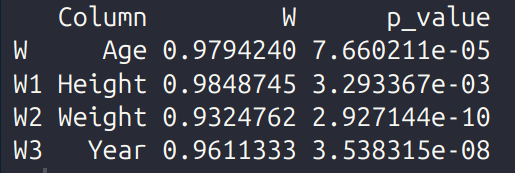
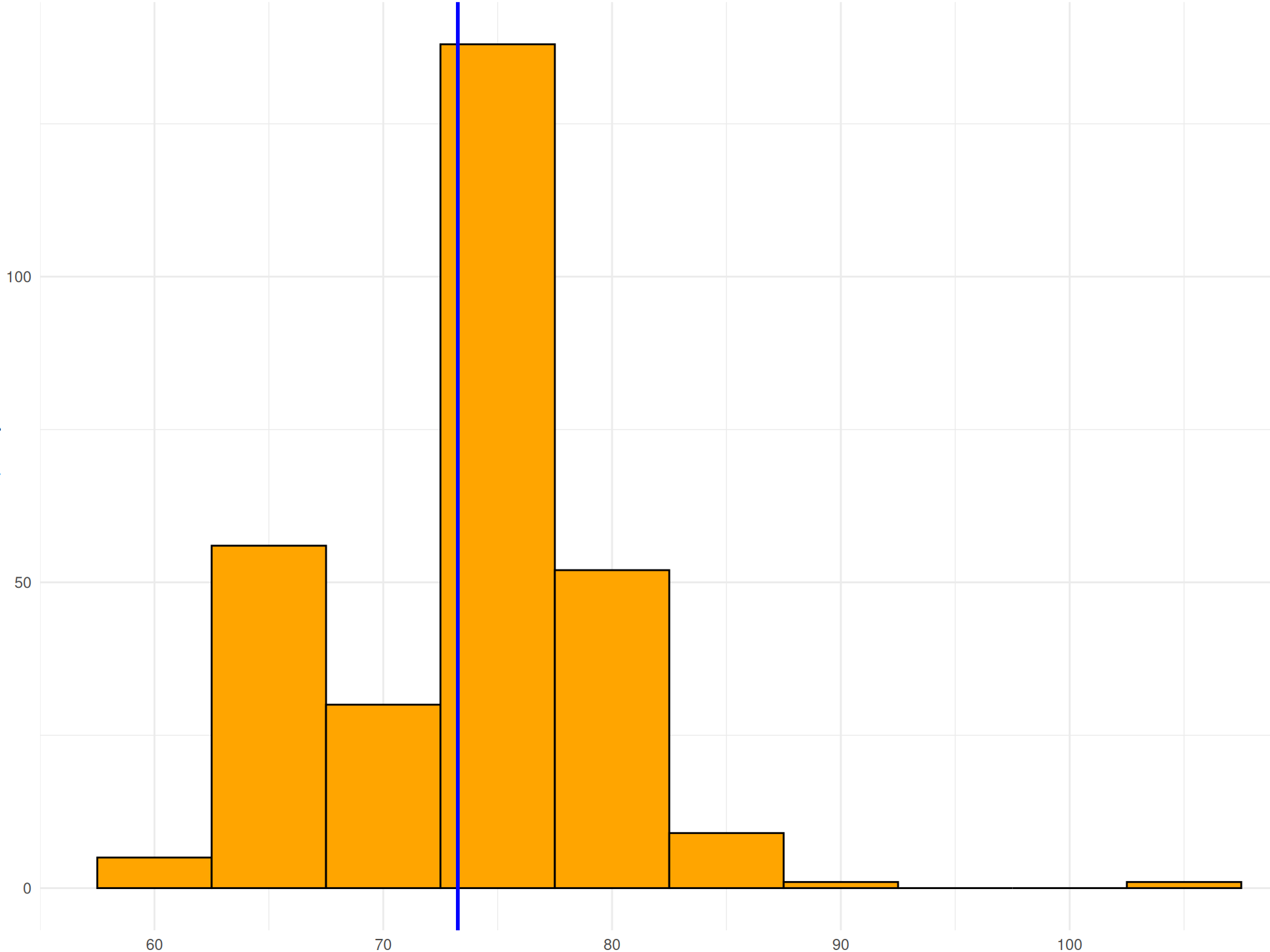
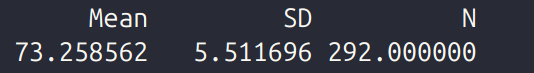


Таблица показывает значения энтропии Шеннона и количества уникальных значений для категориальных переменных в датасете wrestling. Переменная Name имеет наивысшую энтропию (8.106314) и 294 уникальных значения, что указывает на высокую степень разнообразия, так как почти каждый участник имеет уникальное имя, что ожидаемо. Переменные Team и NOC имеют одинаковую энтропию (5.573600) и 64 уникальных значения, что отражает умеренное разнообразие команд и национальных олимпийских комитетов, участвующих в соревнованиях. Переменные Games (энтропия 4.284499, 20 уникальных значений) и City (энтропия 4.124989, 18 уникальных значений) показывают меньшее разнообразие, связанное с ограниченным числом Олимпийских игр и городов проведения, а Medal с энтропией 1.581739 и всего 3 уникальными значениями (Gold, Silver, Bronze) имеет наименьшее разнообрази.

На основе проведенного анализа нормальности и энтропии можно сделать вывод о требуемом тесте. Все числовые переменные (`Age`, `Height`, `Weight`, `Year`) не следуют нормальному распределению, так как p-значения теста Шапиро-Уилка значительно меньше 0.05 , а Q-Q графики показывают ступенчатые узоры и отклонения от линии нормальности, особенно для `Weight` и `Year`. Таблица энтропии подтверждает высокое разнообразие категориальных переменных, таких как `Name` и `Team`, но это не влияет на выбор теста для числовых данных. Поскольку предположения о нормальности не выполняются, для сравнения групп необходимо использовать непараметрический тест, такой как критерий Краскела-Уоллиса, который не требует нормальности и подходит для анализа данных с дискретными или ненормальными распределениями.

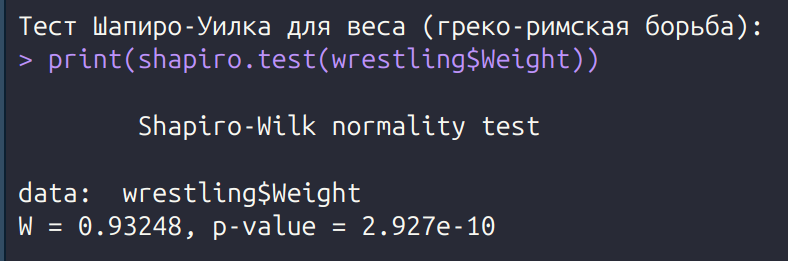
3. Проверьте гипотезу о среднем весе спортсменов выбранного вида спорта.



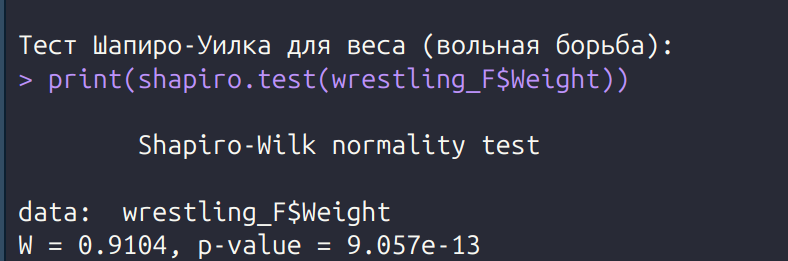
Результаты анализа показали, что средний вес спортсменов в выборке составил 73.26 кг (SD = 5.51), что статистически значимо ниже гипотетического значения 75 кг (z = -5.4, p < 0.001). При размере выборки 292 человека стандартная ошибка среднего оказалась минимальной (0.322), что повышает надежность оценки. Разница в 1.74 кг между выборочным и гипотетическим средним, хотя и значима статистически, имеет малый размер эффекта (Cohen’s d = 0.316), что указывает на ограниченную практическую важность этого различия. Распределение веса близко к нормальному, с пиком частоты в диапазоне 70–80 кг, что подтверждает корректность применения параметрических методов.

4. Проверьте гипотезу о равенстве среднего веса женщин (мужчин) в двух разных выбранных видах спорта (сравнение двух независимых выборок – двухвыборочный критерий.).

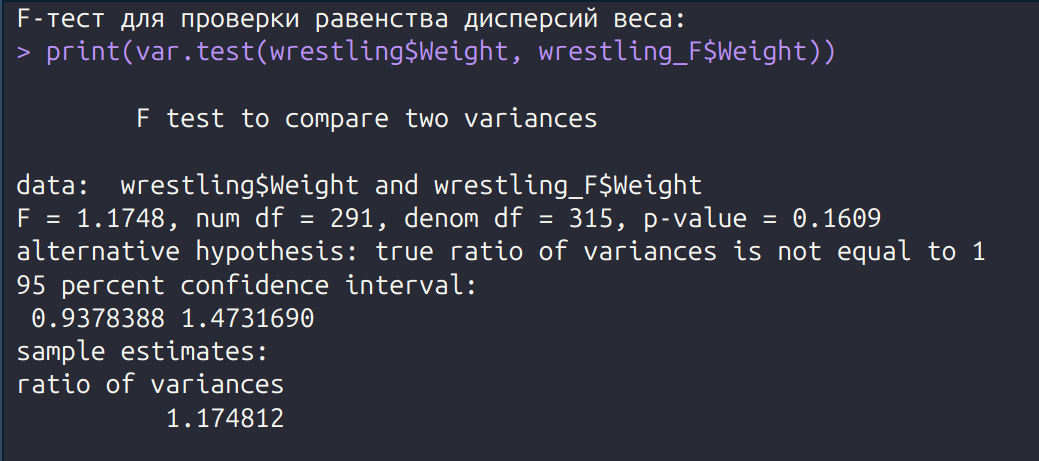
Поскольку p-значение значительно меньше 0.05, мы отвергаем нулевую гипотезу о нормальности распределения веса. Это указывает на то, что распределение веса в данной группе не является нормальным. Такой результат может быть обусловлен наличием выбросов или асимметрией в данных.



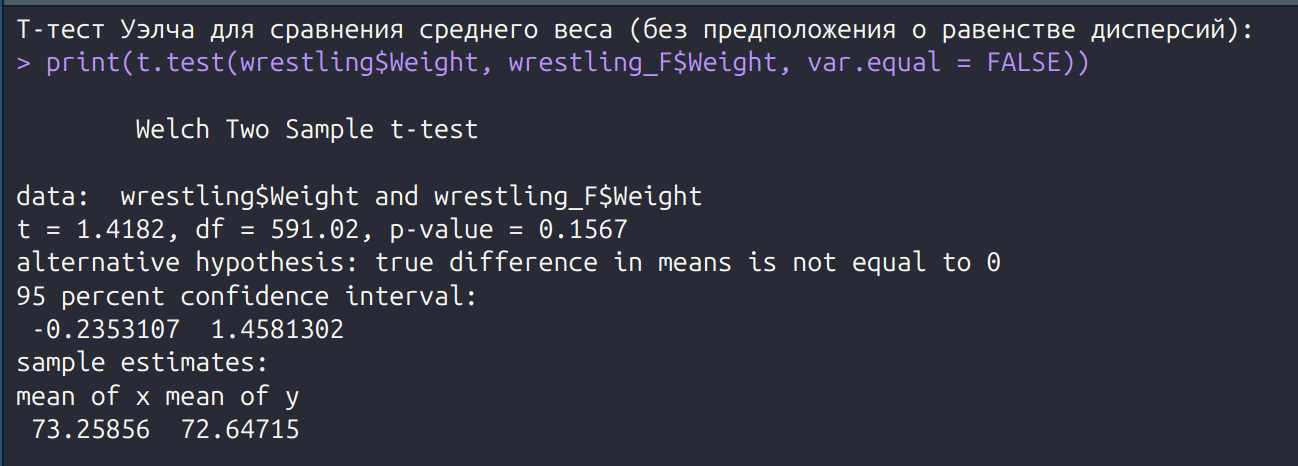
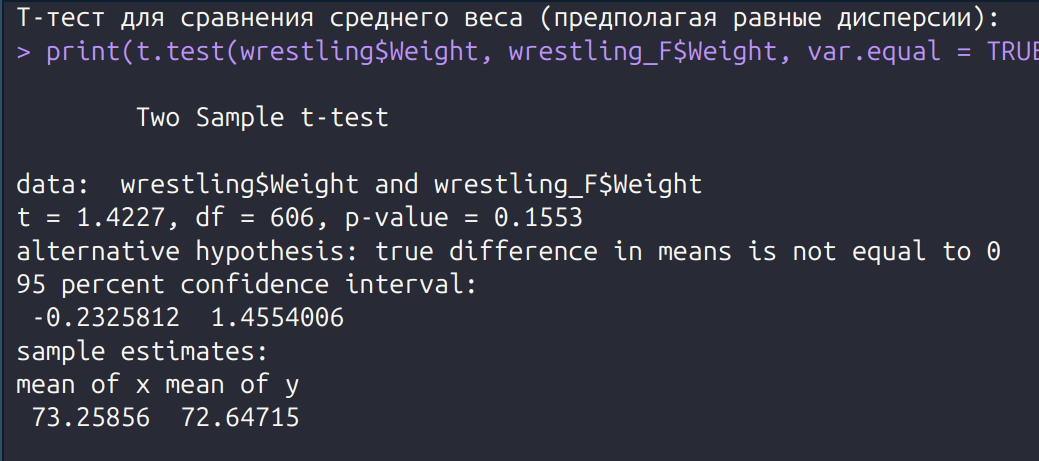
Очень маленькое p-значение (меньше 0.05) также позволяет отвергнуть нулевую гипотезу о нормальном распределении. Это означает, что распределение веса в группе вольной борьбы также не соответствует нормальному закону. Вероятно, данные имеют асимметрию или содержат выбросы, что влияет на результаты теста. Для анализа различий между группами лучше использовать непараметрические методы, избегая предположений о нормальности.



Так как p-значение больше 0.05, мы не можем отвергнуть нулевую гипотезу о равенстве дисперсий. Это означает, что дисперсии веса в обеих группах можно считать статистически равными. Оценка отношения дисперсий составляет 1.1748, а 95%-й доверительный интервал (0.9378, 1.4732) включает 1, что подтверждает вывод о равенстве дисперсий. Для t-теста можно использовать предположение о равных дисперсиях.

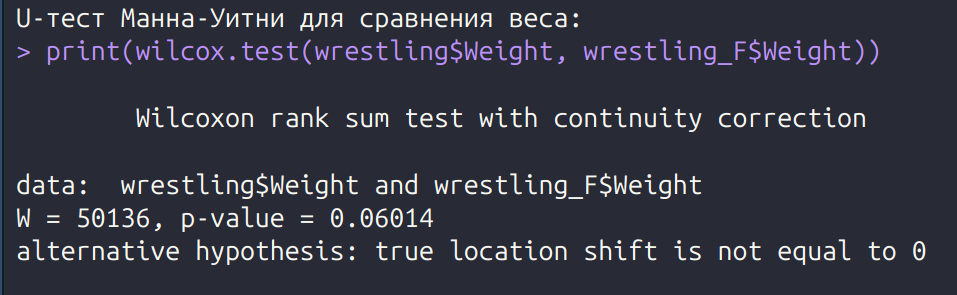


Поскольку p-значение больше 0.05, мы не отвергаем нулевую гипотезу о равенстве средних весов. Средний вес в греко-римской борьбе составляет 73.25856 кг, а в вольной — 72.64715 кг, что указывает на небольшую разницу. Доверительный интервал для разницы средних (-0.2258, 1.4554) включает 0, что подтверждает отсутствие статистически значимых различий. Однако, учитывая ненормальность данных, результаты t-теста следует интерпретировать с осторожностью.



Доверительный интервал для разницы средних (-0.2353, 1.4581) включает 0, подтверждая отсутствие значимых различий. Средние веса остаются теми же: 73.25856 кг для греко-римской борьбы и 72.64715 кг для вольной. Этот тест более устойчив к различиям в дисперсиях, но ненормальность данных делает его менее предпочтительным по сравнению с непараметрическим тестом.

Этот тест не предполагает нормальности данных, что делает его более подходящим в данном случае. Результаты указывают на отсутствие статистически значимых различий в весе между греко-римской и вольной борьбой. Таким образом, U-тест подтверждает выводы t-тестов, но является более надежным в условиях ненормальности.



Распределение веса спортсменов в греко-римской и вольной борьбе не является нормальным, что подтверждают тесты Шапиро-Уилка с p-значениями 2.927×10−10 и 9.057×10−13. Дисперсии веса в обеих группах равны ( p=0.1609), что позволяет использовать t-тест с равными дисперсиями, хотя ненормальность делает U-тест Манна-Уитни предпочтительным. Все тесты показывают отсутствие статистически значимых различий в среднем весе между группами (p>0.05). Средний вес составляет 73.26 кг для греко-римской борьбы и 72.65 кг для вольной, что указывает на минимальную разницу. Таким образом, гипотеза о равенстве среднего веса спортсменов в двух видах борьбы не может быть отвергнута на уровне значимости 0.05.

**Вывод: я** ознакомилась с некоторыми статистическими тестами, принципами их работы. Научилась оценивать нормальность распределения выборки, а также выполнять оценку статистических гипотез.

**Листинг:**

# Подключение библиотек

library(ggplot2)

library(car)

library(stats)

# Чтение данных

data <- read.csv("~/Documents/6 semester/Big data/6laba/data.csv", header = TRUE, sep = ",")

# Фильтрация данных для греко-римской борьбы

wrestling <- data[data$Event == "Wrestling Men's Welterweight, Greco-Roman", ]

wrestling <- wrestling[, !names(wrestling) %in% c("Season", "Sport", "Event", "Sex")]

# Функция для расчета статистик

calculate\_stats <- function(x) {

mean\_val <- mean(x, na.rm = TRUE)

median\_val <- median(x, na.rm = TRUE)

mode\_val <- as.numeric(names(sort(table(x), decreasing = TRUE)[1]))

var\_val <- var(x, na.rm = TRUE)

sd\_val <- sd(x, na.rm = TRUE)

min\_val <- min(x, na.rm = TRUE)

max\_val <- max(x, na.rm = TRUE)

data.frame(

Среднее = mean\_val,

Медиана = median\_val,

Мода = mode\_val,

Дисперсия = var\_val,

Стандартное\_отклонение = sd\_val,

Минимум = min\_val,

Максимум = max\_val

)

}

# Выбор числовых столбцов (исключая ID)

numeric\_cols <- wrestling[, sapply(wrestling, is.numeric) & !names(wrestling) %in% "ID"]

# Расчет статистик для числовых столбцов

stats\_list <- lapply(numeric\_cols, calculate\_stats)

stats\_table <- do.call(rbind, stats\_list)

stats\_table <- cbind(Переменная = names(numeric\_cols), stats\_table)

rownames(stats\_table) <- NULL

# Вывод таблицы статистик

cat("\nТаблица статистических показателей:\n")

print(stats\_table)

# Функция для построения графиков

plot\_analysis <- function(df) {

numeric\_cols <- sapply(df, is.numeric)

numeric\_names <- names(df)[numeric\_cols & !names(df) %in% c("ID", "Year")]

categorical\_names <- names(df)[!numeric\_cols & !names(df) %in% c("Name")]

for (col in numeric\_names) {

p1 <- ggplot(df, aes(y = !!sym(col))) +

geom\_boxplot(fill = "lightblue", width = 0.3) +

ggtitle(paste("Ящик с усами для", col)) +

theme\_minimal() +

theme(

axis.title.x = element\_blank(),

axis.text.x = element\_blank(),

axis.ticks.x = element\_blank()

) +

scale\_y\_continuous(name = col)

p2 <- ggplot(df, aes(x = !!sym(col))) +

geom\_histogram(fill = "orange", bins = 30, color = "black") +

ggtitle(paste("Гистограмма для", col)) +

xlab(col) +

ylab("Частота") +

theme\_minimal()

print(p1)

print(p2)

}

for (col in categorical\_names) {

p <- ggplot(df, aes(x = !!sym(col))) +

geom\_bar(fill = "orange", color = "black") +

ggtitle(paste("Столбчатая диаграмма для", col)) +

xlab(col) +

ylab("Частота") +

theme\_minimal() +

theme(axis.text.x = element\_text(angle = 45, hjust = 1))

print(p)

}

}

# Построение графиков для греко-римской борьбы

cat("\nПостроение графиков для греко-римской борьбы...\n")

plot\_analysis(wrestling)

# Проверка нормальности

normality\_results <- lapply(names(numeric\_cols), function(col) {

shapiro\_result <- shapiro.test(numeric\_cols[[col]])

p <- ggplot(wrestling, aes(sample = !!sym(col))) +

stat\_qq() +

stat\_qq\_line(color = "red") +

ggtitle(paste("Q-Q график для", col)) +

xlab("Теоретические квантили") +

ylab("Выборочные квантили") +

theme\_minimal()

print(p)

data.frame(

Переменная = col,

W = shapiro\_result$statistic,

p\_значение = shapiro\_result$p.value

)

})

# Вывод результатов теста Шапиро-Уилка

normality\_table <- do.call(rbind, normality\_results)

cat("\nРезультаты теста Шапиро-Уилка на нормальность:\n")

print(normality\_table)

# Расчет дисперсии

variance\_results <- lapply(names(numeric\_cols), function(col) {

var\_val <- var(numeric\_cols[[col]], na.rm = TRUE)

data.frame(

Переменная = col,

Дисперсия = var\_val

)

})

# Вывод таблицы дисперсий

variance\_table <- do.call(rbind, variance\_results)

cat("\nДисперсия для числовых переменных:\n")

print(variance\_table)

# Тест Левена (если есть несколько категорий в Medal)

if (length(unique(wrestling$Medal)) > 1) {

levene\_results <- lapply(names(numeric\_cols), function(col) {

test <- leveneTest(wrestling[[col]] ~ wrestling$Medal, data = wrestling)

data.frame(

Переменная = col,

F\_значение = test$`F value`[1],

p\_значение = test$`Pr(>F)`[1]

)

})

levene\_table <- do.call(rbind, levene\_results)

cat("\nТест Левена на гомогенность дисперсий (по медалям):\n")

print(levene\_table)

} else {

cat("\nТест Левена пропущен: в столбце Medal только одна категория или данные отсутствуют.\n")

}

# Расчет энтропии для категориальных переменных

calculate\_entropy <- function(x) {

x <- x[!is.na(x)]

if (length(x) == 0) return(NA)

probs <- table(x) / length(x)

entropy <- -sum(probs \* log2(probs), na.rm = TRUE)

return(entropy)

}

# Выбор категориальных столбцов

categorical\_cols <- wrestling[, !sapply(wrestling, is.numeric)]

# Расчет энтропии

entropy\_results <- lapply(names(categorical\_cols), function(col) {

entropy\_val <- calculate\_entropy(categorical\_cols[[col]])

unique\_count <- length(unique(categorical\_cols[[col]][!is.na(categorical\_cols[[col]])]))

data.frame(

Переменная = col,

Шенноновская\_энтропия = entropy\_val,

Уникальные\_значения = unique\_count

)

})

# Вывод таблицы энтропии

entropy\_table <- do.call(rbind, entropy\_results)

cat("\nШенноновская энтропия и количество уникальных значений для категориальных переменных:\n")

print(entropy\_table)

# Тест Уилкоксона для веса (греко-римская борьба)

wilcox\_test <- wilcox.test(wrestling$Weight, mu = 75, alternative = "two.sided", conf.int = TRUE)

cat("\nТест Уилкоксона для веса (гипотеза: среднее = 75 кг):\n")

print(wilcox\_test)

# Описательные статистики для веса

weight\_summary <- c(

Среднее = mean(wrestling$Weight, na.rm = TRUE),

Стандартное\_отклонение = sd(wrestling$Weight, na.rm = TRUE),

N = length(na.omit(wrestling$Weight))

)

cat("\nОписательные статистики для веса:\n")

print(weight\_summary)

# Гистограмма веса

p1 <- ggplot(wrestling, aes(x = Weight)) +

geom\_histogram(binwidth = 5, fill = "orange", color = "black") +

geom\_vline(xintercept = mean(wrestling$Weight, na.rm = TRUE), color = "blue", linetype = "solid", size = 1) +

ggtitle("Гистограмма веса с указанием среднего и гипотетического значения (75 кг)") +

xlab("Вес (кг)") +

ylab("Частота") +

theme\_minimal()

print(p1)

# Фильтрация данных для вольной борьбы (исправлено)

wrestling\_F <- data[data$Event == "Wrestling Men's Welterweight, Freestyle", ]

wrestling\_F <- wrestling\_F[, !names(wrestling\_F) %in% c("Season", "Sport", "Event", "Sex")]

# Проверка нормальности для веса

cat("\nТест Шапиро-Уилка для веса (греко-римская борьба):\n")

print(shapiro.test(wrestling$Weight))

cat("\nТест Шапиро-Уилка для веса (вольная борьба):\n")

print(shapiro.test(wrestling\_F$Weight))

# Проверка равенства дисперсий

cat("\nF-тест для проверки равенства дисперсий веса:\n")

print(var.test(wrestling$Weight, wrestling\_F$Weight))

# t-тест (если нормальность и равенство дисперсий подтверждены)

cat("\nT-тест для сравнения среднего веса (предполагая равные дисперсии):\n")

print(t.test(wrestling$Weight, wrestling\_F$Weight, var.equal = TRUE))

# t-тест Уэлча (если дисперсии не равны)

cat("\nT-тест Уэлча для сравнения среднего веса (без предположения о равенстве дисперсий):\n")

print(t.test(wrestling$Weight, wrestling\_F$Weight, var.equal = FALSE))

# U-тест Манна-Уитни (если данные не нормальны)

cat("\nU-тест Манна-Уитни для сравнения веса:\n")

print(wilcox.test(wrestling$Weight, wrestling\_F$Weight))