Module 2: Assignment

Aaron Hum

Stephanie Moyerman, PhD

LSC 541: Statistics for Biological Data Science I

2024-07-12

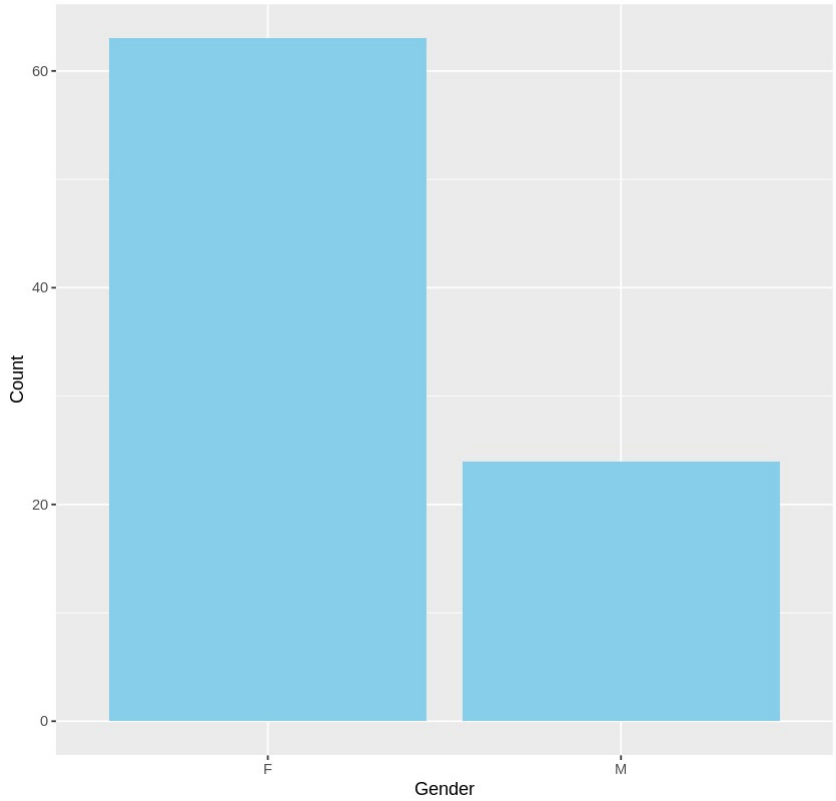
```r
# load library
library(ggplot2)

# read in data
data <- read.table('data1_LSC598.txt', header = T)
data
```

```
   age_month gender vitD_level group
1  60        F      47.5       1
2  50        M      39.0       1
3  35        F      13.6       1
4  50        F      16.7       1
5  61        F      32.7       0
6  55        F      23.2       0
7  54        F      36.4       0
8  60        F      38.1       1
9  47        F      28.2       1
10 65        M      34.1       1
11 22        F      25.7       1
12 52        M      39.9       1
13 50        F      45.2       1
14 51        F      47.8       0
15 33        F      42.4       0
16 60        F      39.3       1
17 65        F      23.4       1
18 74        F      36.6       0
19 51        F      18.4       1
20 57        F      87.2       0
21 51        F      35.1       0
22 54        F      38.5       1
23 53        F      28.2       1
24 56        F      30.6       1
25 47        F      52.5       0
26 46        M      23.3       0
27 41        M      45.4       1
28 49        F      38.6       0
29 66        M      46.0       1
```

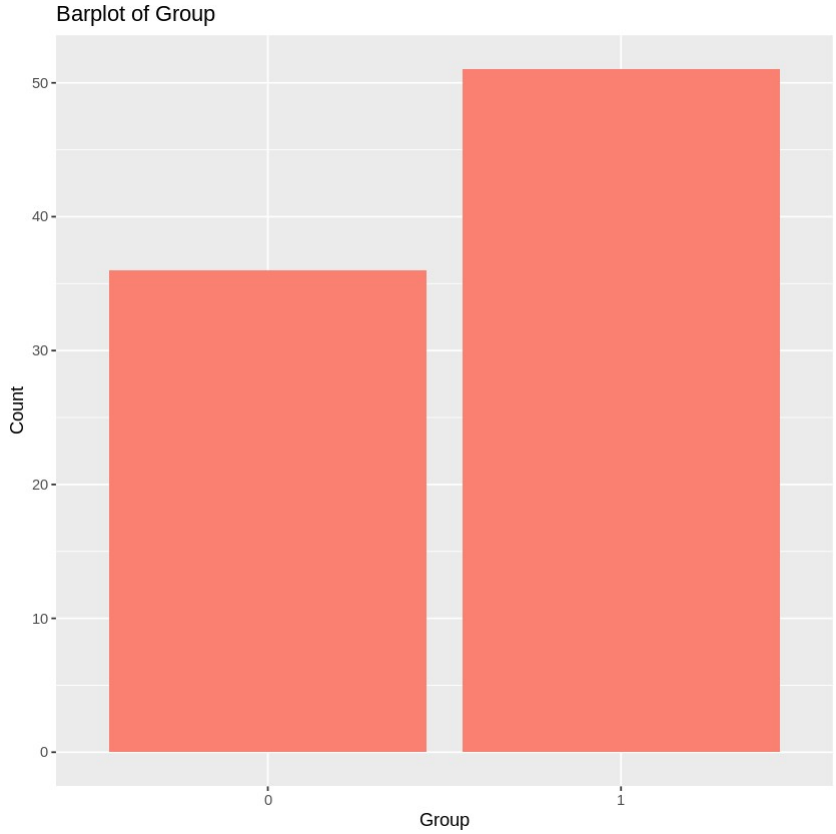
```
30 48        M      51.3       1
⋮  ⋮         ⋮      ⋮          ⋮
58 57        F      24.9       1
59 40        M      45.5       1
60 57        M      28.6       0
61 27        F      63.6       1
62 45        F      21.1       0
63 27        F      19.2       1
64 51        M      24.0       0
65 50        M      50.1       0
66 50        F      26.1       1
67 64        F      16.1       1
68 52        M      34.7       1
69 56        M      16.0       0
70 63        M      26.8       0
71 53        M      37.4       1
72 38        F      39.7       0
73 40        M      44.8       0
74 49        F        NA       0
75 54        F      31.3       0
76 25        F      32.7       1
77 63        F      31.5       1
78 50        M      40.6       1
79 64        M      48.0       0
80 51        F      39.5       1
81 49        F      46.3       1
82 57        F      31.3       1
83 50        F      27.0       1
84 71        F      46.6       0
85 51        F      50.9       1
86 48        M      31.3       0
87 47        F      15.1       0
```

```r
# 1. barplots for qualitative variables

# barplot for 'gender'
ggplot(data, aes(x = gender)) +
  geom_bar(fill = "skyblue") +
  labs(title = "Barplot of Gender", x = "Gender", y = "Count")

# barplot for 'group'
ggplot(data, aes(x = factor(group))) +
  geom_bar(fill = "salmon") +
  labs(title = "Barplot of Group", x = "Group", y = "Count")
```

Barplot of Gender

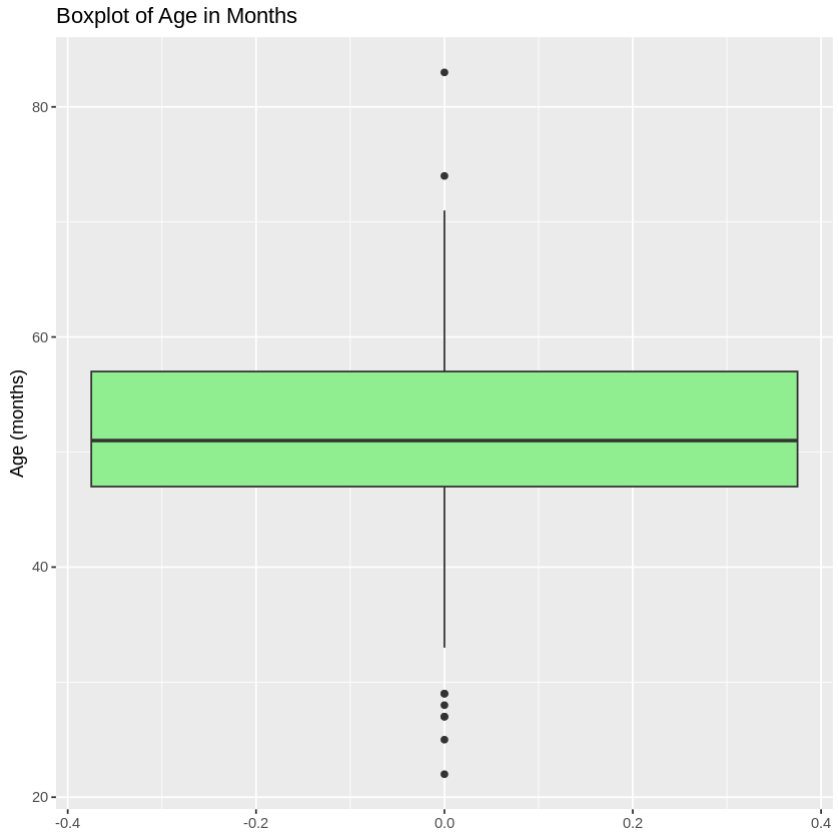


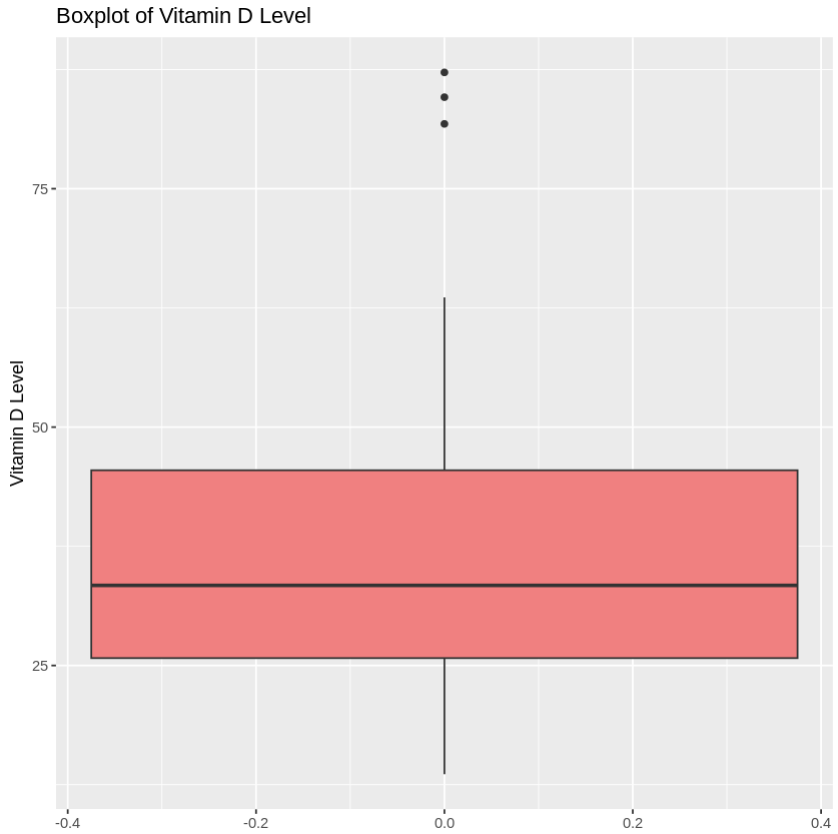
Barplot of Group



```r
# 2. boxplots for quantitative variables

# boxplot for 'age_month'
ggplot(data, aes(y = age_month)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "Boxplot of Age in Months", y = "Age (months)")

# boxplot for 'vitD_level'
ggplot(data, aes(y = vitD_level)) +
  geom_boxplot(fill = "lightcoral") +
  labs(title = "Boxplot of Vitamin D Level", y = "Vitamin D Level")
```

```
Warning message:
“Removed 1 row containing non-finite outside the scale range
(`stat_boxplot()`).”
```






```r
# 3. one-sample hypothesis testing for each continuous variable

# one-sample t-test for 'age_month'
t_test_age_month <- t.test(data$age_month, mu = 0)
print(t_test_age_month)

# one-sample t-test for 'vitD_level'
t_test_vitD_level <- t.test(data$vitD_level, mu = 0)
print(t_test_vitD_level)
```

```
	One Sample t-test

data:  data$age_month
t = 43.792, df = 86, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 48.72877 53.36318
sample estimates:
mean of x
 51.04598


	One Sample t-test
```

```
data:  data$vitD_level
t = 21.418, df = 85, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 32.62538 39.30253
sample estimates:
mean of x
 35.96395
```

Conclusion of t-tests

For both age_month and vitD_level, the p-values are extremely small, and the confidence intervals do not include 0. Therefore, we reject the null hypothesis in both cases. This means that there is strong evidence that the true means of age_month and vitD_level are not equal to 0. The average age of the subjects is approximately 51 months, and the average vitamin D level is approximately 36.