cancer_Q2

Yunya He

2024-10-08

Read the Excel file

Modify according to actual file path

```
data <- read_excel("Cancer.xlsx")
```

```
# Check the structure of the data
str(data)
```

tibble [100 x 4] (S3: tbl_df/tbl/data.frame) $ Type : chr [1:100] “Stomach” “Stomach” “Stomach” “Stomach” ... $ Sex : chr [1:100] “F” “M” “F” “F” ... $ Age : num [1:100] 61 69 62 66 42 79 76 54 62 69 ... $ Survival: num [1:100] 121 12 9 18 258 43 142 36 149 182 ...

1. ANOVA model

First, we constructed an ANOVA model to test whether there are significant differences in the mean ages of patients across different cancer types.

```
anova_model <- aov(Age ~ Type, data = data)
anova_summary <- summary(anova_model)
anova_summary
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
```

Type 21 3682 175.3 1.499 0.103 Residuals 78 9126 117.0

In this result, Df represents degrees of freedom, Sum Sq is the sum of squares, Mean Sq is the mean square, the F value is 1.499, and the p-value is 0.103. Since the p-value is greater than 0.05, we cannot reject the null hypothesis, indicating that there is no significant difference in the mean ages among different cancer types.

2. Normality test

Next, we conducted a normality test to verify whether the residuals of the ANOVA model follow a normal distribution.

```
shapiro_test <- shapiro.test(residuals(anova_model))
shapiro_test
```

```
Shapiro-Wilk normality test
```

data: residuals(anova_model) W = 0.97823, p-value = 0.0967

The p-value of 0.0967 is greater than 0.05, indicating that we do not have enough evidence to reject the normality assumption for the residuals, suggesting they approximately follow a normal distribution.

3. Homogeneity of variance test

We also performed a homogeneity of variance test using Levene's test to confirm whether the variances across different cancer types are equal. Ensure Type is treated as a factor

```
data$Type <- as.factor(data$Type)
levene_test <- leveneTest(Age ~ Type, data = data)
levene_test
```

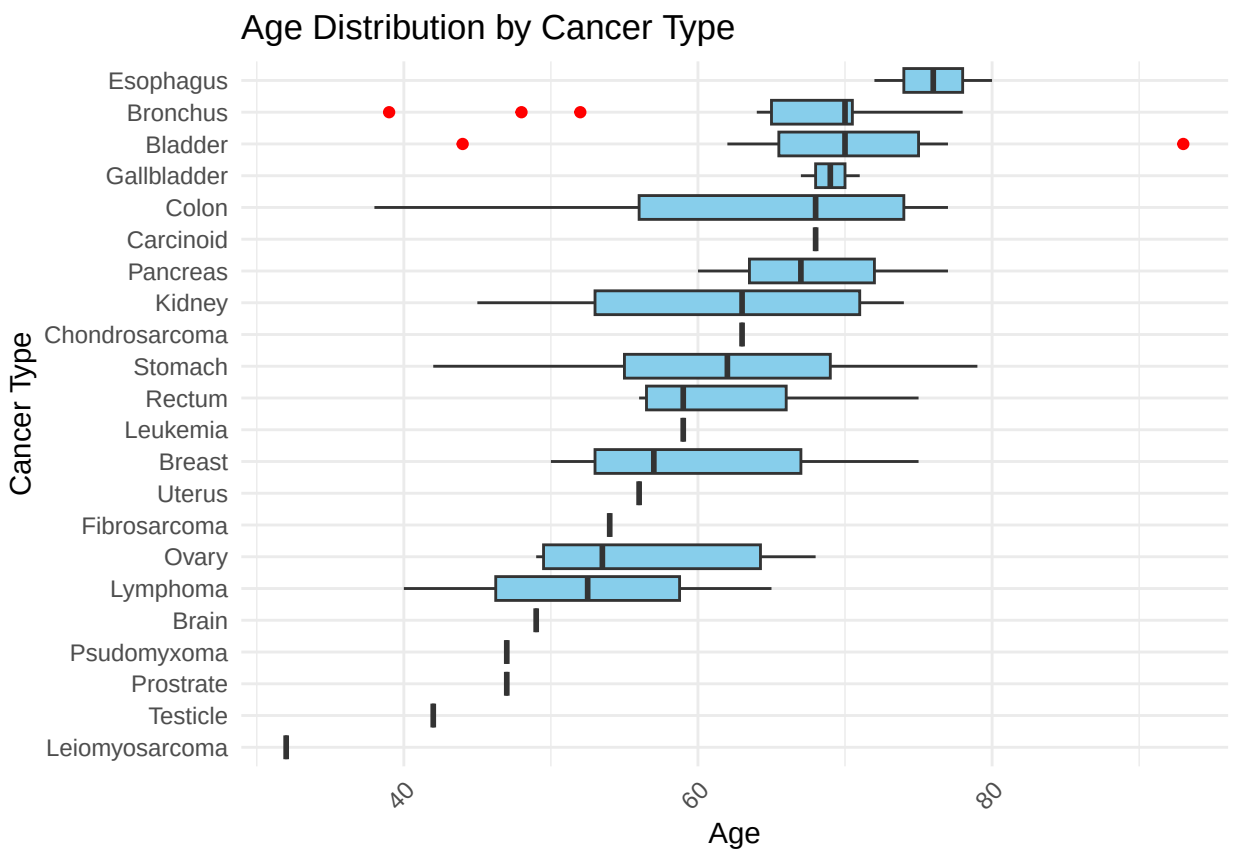
Levene's Test for Homogeneity of Variance (center = median) Df F value Pr(>F) group 21 0.7349 0.7844 78

The p-value of 0.7844 is greater than 0.05, indicating that we do not have enough evidence to reject the homogeneity of variance assumption, suggesting that the variances are relatively equal across groups.

4.Visualize age distribution - Boxplot

Create a horizontal boxplot

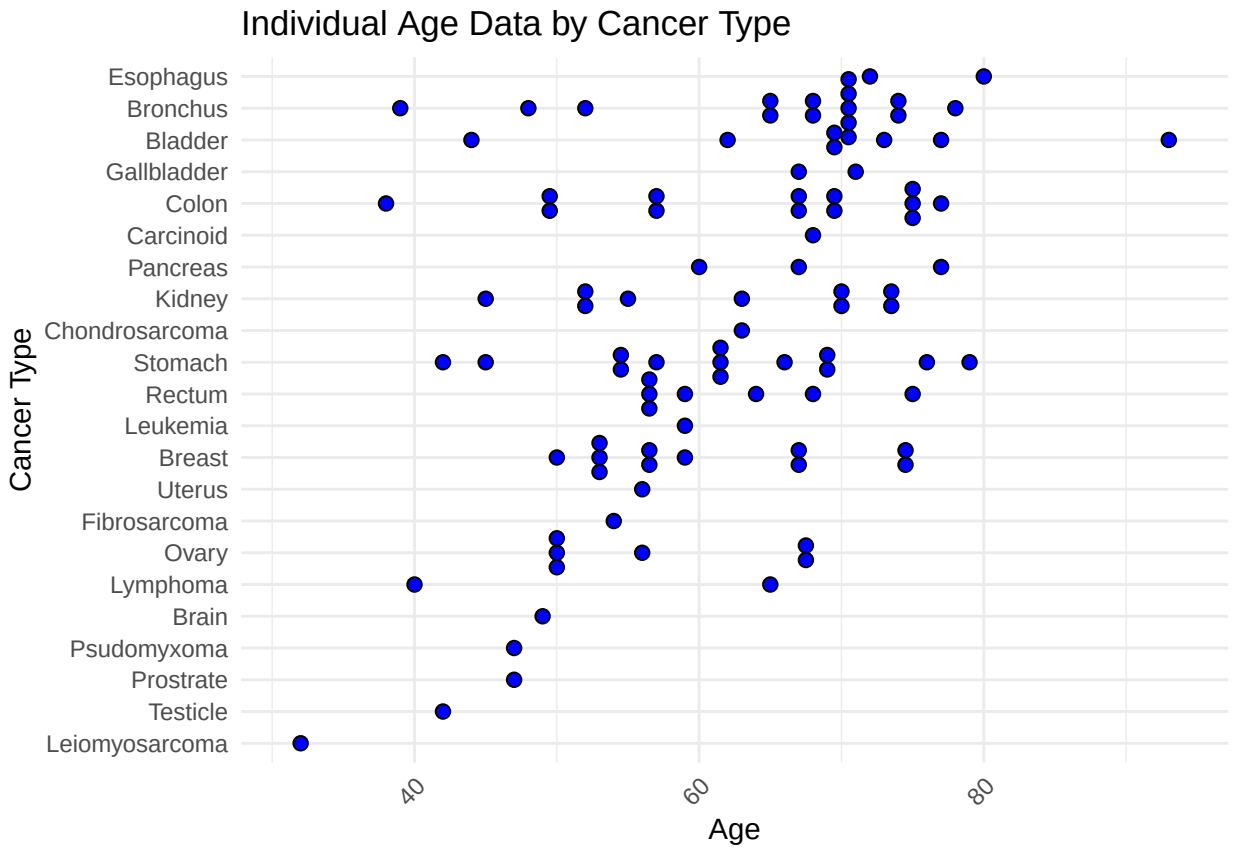
```
ggplot(data, aes(x = reorder(Type, Age, FUN = median), y = Age)) +
  geom_boxplot(fill = "skyblue", outlier.color = "red") +
  coord_flip() +  # Flip coordinates for horizontal display
  labs(title = "Age Distribution by Cancer Type",
       x = "Cancer Type",
       y = "Age") +
  theme_minimal() +  # Use a minimal theme for better aesthetics
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Adjust x-axis text for better readability
```



Create a dot plot

```
ggplot(data, aes(x = reorder(Type, Age, FUN = median), y = Age)) +
  geom_dotplot(binaxis = 'y', stackdir = 'center', fill = "blue", dotsize = 0.5) +
  coord_flip() +  # Flip coordinates for horizontal display
  labs(title = "Individual Age Data by Cancer Type",
       x = "Cancer Type",
       y = "Age") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Bin width defaults to 1/30 of the range of the data. Pick better value with
## `binwidth`.
```


Individual Age Data by Cancer Type
Esophagus
Bronchus
Bladder
Gallbladder
Colon
Carcinoid
Pancreas
Kidney
Chondrosarcoma
Stomach
Rectum
Leukemia
Breast
Uterus
Fibrosarcoma
Ovary
Lymphoma
Brain
Psudomyxoma
Prostrate
Testicle
Leiomyosarcoma
Cancer Type
40
60
80
Age