STAT631 Mini Project Report

David Teng

2025-05-13

Data Description

This analysis uses the Right Heart Catheterization (RHC) dataset, which includes 5,735 critically ill adult patients from the SUPPORT study (Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments), conducted from 1989 to 1994 across five U.S. teaching hospitals. The dataset is publicly available through the Vanderbilt University Department of Biostatistics. For the purpose of this study, the outcome of interest is hospital length of stay, defined as the number of days from admission to discharge. The independent variables examined are sex (male or female), age group (categorized as <50, 50–65, 65–80, and 80+), and history of cardiovascular disease (yes or no).

Right Heart Catheterization (RHC) Dataset Overview

Source: Vanderbilt University Department of Biostatistics

Direct Download: https://hbiostat.org/data/repo/rhc.csv

Dataset Documentation: https://search.r-project.org/CRAN/refmans/ATbounds/html/RHC.html

Study Reference: Connors, A.F., et al. (1996). "The effectiveness of right heart catheterization in the initial care of critically ill patients." JAMA, 276(11), 889–897. DOI: 10.1001/jama.1996.03540110043030

Research Question

This study examines whether sex, age group, and history of cardiovascular disease (cardiohx) individually or interactively influence hospital length of stay among critically ill patients.

Load Libraries and Data

```r
library(ggplot2)
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
library(multcomp)
```

```
Loading required package: mvtnorm

Loading required package: survival

Loading required package: TH.data

Loading required package: MASS


Attaching package: 'MASS'

The following object is masked from 'package:dplyr':

    select


Attaching package: 'TH.data'

The following object is masked from 'package:MASS':

    geyser
```

```r
library(stats)
library(car)
```

```
Loading required package: carData
```

```
Attaching package: 'car'
```

```
The following object is masked from 'package:dplyr':

    recode
```

```r
df <- read.csv("rhc.csv")
summary(df)
```

```
       X                cat1               cat2                 ca
 Min.   :   1    Length:5735        Length:5735         Length:5735
 1st Qu.:1434    Class :character    Class :character    Class :character
 Median :2868    Mode  :character    Mode  :character    Mode  :character
 Mean   :2868
 3rd Qu.:4302
 Max.   :5735

    sadmdte          dschdte          dthdte           lstctdte
 Min.   :10754    Min.   :10757    Min.   :10757    Min.   :10756
 1st Qu.:11164    1st Qu.:11184    1st Qu.:11267    1st Qu.:11316
 Median :11759    Median :11777    Median :11832    Median :11868
 Mean   :11639    Mean   :11660    Mean   :11754    Mean   :11781
 3rd Qu.:12097    3rd Qu.:12120    3rd Qu.:12208    3rd Qu.:12244
 Max.   :12441    Max.   :12560    Max.   :12783    Max.   :12644
                  NA's   :1        NA's   :2013
    death             cardiohx           chfhx             dementhx
 Length:5735       Min.   :0.0000    Min.   :0.000    Min.   :0.00000
 Class :character   1st Qu.:0.0000    1st Qu.:0.000    1st Qu.:0.00000
 Mode  :character   Median :0.0000    Median :0.000    Median :0.00000
                    Mean   :0.1766    Mean   :0.178    Mean   :0.09834
                    3rd Qu.:0.0000    3rd Qu.:0.000    3rd Qu.:0.00000
                    Max.   :1.0000    Max.   :1.000    Max.   :1.00000

    psychhx            chrpulhx          renalhx            liverhx
```

```
Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000
Median :0.00000   Median :0.0000   Median :0.00000   Median :0.00000
Mean   :0.06731   Mean   :0.1899   Mean   :0.04446   Mean   :0.06992
3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000
Max.   :1.00000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000

   gibledhx          malighx           immunhx           transhx
Min.   :0.00000   Min.   :0.0000   Min.   :0.000    Min.   :0.0000
1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.000    1st Qu.:0.0000
Median :0.00000   Median :0.0000   Median :0.000    Median :0.0000
Mean   :0.03226   Mean   :0.2295   Mean   :0.269    Mean   :0.1154
3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:1.000    3rd Qu.:0.0000
Max.   :1.00000   Max.   :1.0000   Max.   :1.000    Max.   :1.0000

    amihx             age              sex                edu
Min.   :0.00000   Min.   : 18.04   Length:5735       Min.   : 0.00
1st Qu.:0.00000   1st Qu.: 50.15   Class :character  1st Qu.:10.00
Median :0.00000   Median : 64.05   Mode  :character  Median :12.00
Mean   :0.03487   Mean   : 61.38                     Mean   :11.68
3rd Qu.:0.00000   3rd Qu.: 73.93                     3rd Qu.:13.00
Max.   :1.00000   Max.   :101.85                     Max.   :30.00

   surv2md1          das2d3pc          t3d30            dth30
Min.   :0.0000   Min.   :11.00    Min.   : 2.00    Length:5735
1st Qu.:0.4709   1st Qu.:16.06    1st Qu.:16.00    Class :character
Median :0.6280   Median :19.75    Median :30.00    Mode  :character
Mean   :0.5925   Mean   :20.50    Mean   :23.61
3rd Qu.:0.7430   3rd Qu.:23.43    3rd Qu.:30.00
Max.   :0.9620   Max.   :33.00    Max.   :30.00

     aps1             scoma1           meanbp1           wblc1
Min.   :  3.00   Min.   :  0      Min.   :  0.00   Min.   :  0.000
1st Qu.: 41.00   1st Qu.:  0      1st Qu.: 50.00   1st Qu.:  8.398
Median : 54.00   Median :  0      Median : 63.00   Median : 14.100
Mean   : 54.67   Mean   : 21      Mean   : 78.52   Mean   : 15.645
3rd Qu.: 67.00   3rd Qu.: 41      3rd Qu.:115.00   3rd Qu.: 20.049
Max.   :147.00   Max.   :100      Max.   :259.00   Max.   :192.000

     hrt1             resp1            temp1            pafi1
Min.   :  0.0    Min.   :  0.00   Min.   :27.00    Min.   : 11.6
1st Qu.: 97.0    1st Qu.: 14.00   1st Qu.:36.09    1st Qu.:133.3
Median :124.0    Median : 30.00   Median :38.09    Median :202.5
```

```
Mean   :115.2   Mean   : 28.09   Mean   :37.62   Mean   :222.3
3rd Qu.:141.0   3rd Qu.: 38.00   3rd Qu.:39.00   3rd Qu.:316.6
Max.   :250.0   Max.   :100.00   Max.   :43.00   Max.   :937.5

     alb1             hema1            bili1              crea1
Min.   : 0.300   Min.   : 2.00   Min.   : 0.09999   Min.   : 0.09999
1st Qu.: 2.600   1st Qu.:26.10   1st Qu.: 0.79993   1st Qu.: 1.00000
Median : 3.500   Median :30.00   Median : 1.00977   Median : 1.50000
Mean   : 3.093   Mean   :31.87   Mean   : 2.26707   Mean   : 2.13302
3rd Qu.: 3.500   3rd Qu.:36.30   3rd Qu.: 1.39990   3rd Qu.: 2.39990
Max.   :29.000   Max.   :66.19   Max.   :58.19531   Max.   :25.09766

     sod1             pot1            paco21             ph1
Min.   :101.0   Min.   : 1.100   Min.   :  1.00   Min.   :6.579
1st Qu.:132.0   1st Qu.: 3.400   1st Qu.: 31.00   1st Qu.:7.340
Median :136.0   Median : 3.800   Median : 37.00   Median :7.400
Mean   :136.8   Mean   : 4.067   Mean   : 38.75   Mean   :7.388
3rd Qu.:142.0   3rd Qu.: 4.600   3rd Qu.: 42.00   3rd Qu.:7.460
Max.   :178.0   Max.   :11.898   Max.   :156.00   Max.   :7.770

    swang1            wtkilo1            dnr1             ninsclas
Length:5735      Min.   :  0.00   Length:5735      Length:5735
Class :character 1st Qu.: 56.30   Class :character Class :character
Mode  :character Median : 70.00   Mode  :character Mode  :character
                 Mean   : 67.83
                 3rd Qu.: 83.70
                 Max.   :244.00

     resp             card             neuro            gastr
Length:5735      Length:5735      Length:5735      Length:5735
Class :character Class :character Class :character Class :character
Mode  :character Mode  :character Mode  :character Mode  :character




    renal            meta             hema             seps
Length:5735      Length:5735      Length:5735      Length:5735
Class :character Class :character Class :character Class :character
Mode  :character Mode  :character Mode  :character Mode  :character
```

```
     trauma                ortho                   adld3p                 urin1
 Length:5735          Length:5735          Min.    :0.000       Min.    :   0
 Class :character     Class :character     1st Qu.:0.000        1st Qu.:1110
 Mode  :character     Mode  :character     Median :0.000        Median :1927
                                           Mean    :1.182       Mean    :2192
                                           3rd Qu.:2.000        3rd Qu.:2955
                                           Max.    :7.000       Max.    :9000
                                           NA's    :4296        NA's    :3028
      race                income                  ptid
 Length:5735          Length:5735          Min.    :    5
 Class :character     Class :character     1st Qu.: 2562
 Mode  :character     Mode  :character     Median : 5131
                                           Mean    : 5134
                                           3rd Qu.: 7689
                                           Max.    :10278
```

Data Preparation

```r
# Create age group categories from continuous age variable
df$age_group <- cut(df$age,
                    breaks = c(0, 50, 65, 80, 100),
                    labels = c("<50", "50-65", "65-80", "80+"),
                    right = FALSE)

# Convert categorical variables to factors
df$sex <- as.factor(df$sex)
df$cardiohx <- as.factor(df$cardiohx)

# Ensure death is a factor and create numeric binary version (if not already)
df$death <- as.factor(df$death)
df$death_num <- ifelse(df$death == "Yes", 1, 0)

# Select variables and drop rows with any missing values in selected columns
df_clean <- na.omit(df[, c("death", "sex", "age", "age_group", "cardiohx",
                           "death_num", "race", "income", "adld3p",
                           "urin1", "dschdte", "sadmdte")])

# Ensure date columns are in Date or DateTime format
df_clean$dschdte <- as.Date(df_clean$dschdte)
```

```
df_clean$sadmdte <- as.Date(df_clean$sadmdte)

# Calculate length of stay in days
df_clean$length_of_stay <- as.numeric(df_clean$dschdte - df_clean$sadmdte)

summary(df_clean)
```

```
 death        sex              age          age_group   cardiohx   death_num
 No :365   Female:275   Min.   :18.75   <50   :184   0:471   Min.   :0.0000
 Yes:269   Male  :359   1st Qu.:47.79   50-65:174   1:163   1st Qu.:0.0000
                        Median :63.17   65-80:206           Median :0.0000
                        Mean   :60.03   80+  : 70           Mean   :0.4243
                        3rd Qu.:72.59                       3rd Qu.:1.0000
                        Max.   :95.40                       Max.   :1.0000
     race               income            adld3p           urin1
 Length:634         Length:634         Min.   :0.000   Min.   :   0
 Class :character   Class :character   1st Qu.:0.000   1st Qu.:1335
 Mode  :character   Mode  :character   Median :0.000   Median :2228
                                       Mean   :1.058   Mean   :2451
                                       3rd Qu.:1.000   3rd Qu.:3278
                                       Max.   :7.000   Max.   :9000
    dschdte               sadmdte            length_of_stay
 Min.   :2002-01-10   Min.   :2002-01-07   Min.   :  2.00
 1st Qu.:2002-07-31   1st Qu.:2002-07-15   1st Qu.:  7.00
 Median :2003-01-01   Median :2002-12-14   Median : 11.00
 Mean   :2003-01-13   Mean   :2002-12-30   Mean   : 14.89
 3rd Qu.:2003-07-06   3rd Qu.:2003-06-13   3rd Qu.: 17.00
 Max.   :2004-03-05   Max.   :2004-01-24   Max.   :107.00
```

Assumption Checking Before ANOVA

1. **Independence**

Assumed by study design (random sampling or independent subjects).
No formal test—assume valid if no clustering/repeated measures.

2. **Normality of Residuals**

```
# Fit the model:
anova_model <- aov(length_of_stay ~ sex * age_group * cardiohx, data = df_clean)
# Check residual normality:
```
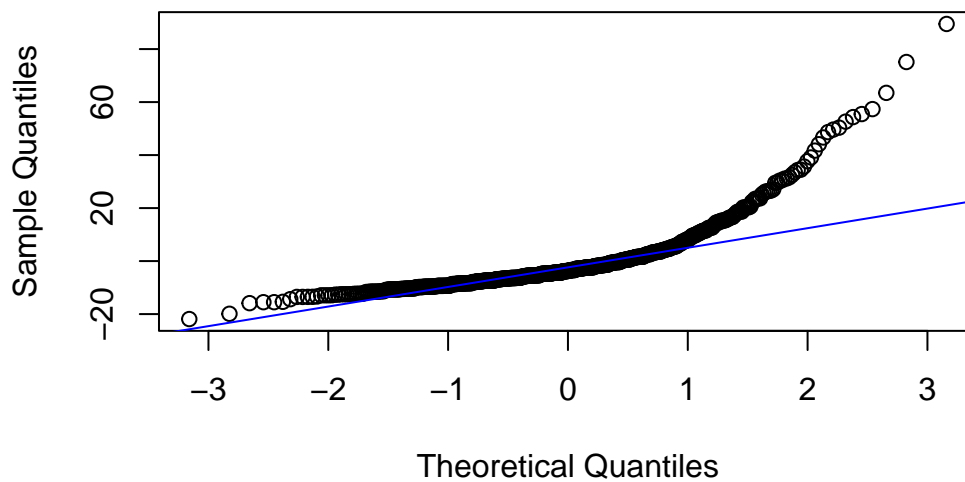
```
resid_anova <- residuals(anova_model)
shapiro.test(resid_anova)
```

```
    Shapiro-Wilk normality test

data:  resid_anova
W = 0.76042, p-value < 2.2e-16
```

```
qqnorm(resid_anova)
qqline(resid_anova, col = "blue")
```

Normal Q–Q Plot

Since $p < 0.05$ **Normality of Residuals** is violated $\rightarrow$ Consider **transformation (e.g., log, sqrt) or** Logistic Regression

3. **Equal Variance**

```
# Levene's Test
library(car)
leveneTest(resid_anova ~ interaction(sex, age_group, cardiohx), data = df_clean)
```

```
Levene's Test for Homogeneity of Variance (center = median)
        Df F value Pr(>F)
group   15  1.3023 0.1948
       618
```

```
# Fit the ANOVA model
anova_model <- aov(length_of_stay ~ sex * age_group * cardiohx, data = df_clean)

# Extract residuals and fitted values
resid_anova <- residuals(anova_model)
fitted_anova <- fitted(anova_model)

# Plot: Residuals vs. Fitted Values
plot(fitted_anova, resid_anova,
     xlab = "Fitted Values",
     ylab = "Residuals",
     main = "Residuals vs. Fitted Values",
     pch = 20,
     col = "steelblue")
abline(h = 0, col = "red", lwd = 2)
```

Residuals vs. Fitted Values
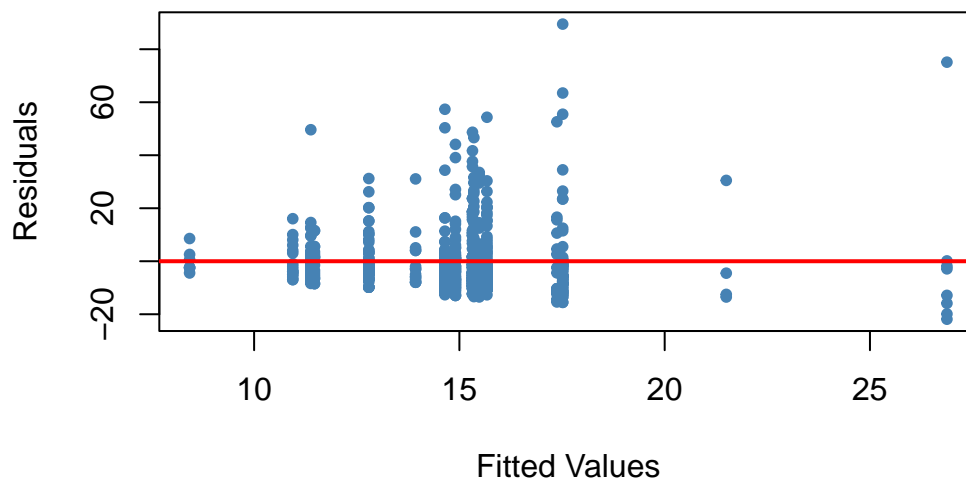
Interpretation:

Levene $p > 0.05 \rightarrow$ Variances are **equal**

Plot: Look for random scatter around 0 (no funnel pattern)

Option 1: Log Transformation

```
# Add a small constant to avoid log(0), if needed
df_clean$log_los <- log(df_clean$length_of_stay + 1)

# Fit the model using log-transformed outcome
anova_model_log <- aov(log_los ~ sex * age_group * cardiohx, data = df_clean)

# Check residuals again
resid_log <- residuals(anova_model_log)
shapiro.test(resid_log)
```

```
    Shapiro-Wilk normality test

data:  resid_log
W = 0.98635, p-value = 1.199e-05
```

```
qqnorm(resid_log)
qqline(resid_log, col = "blue")
```

Normal Q–Q Plot
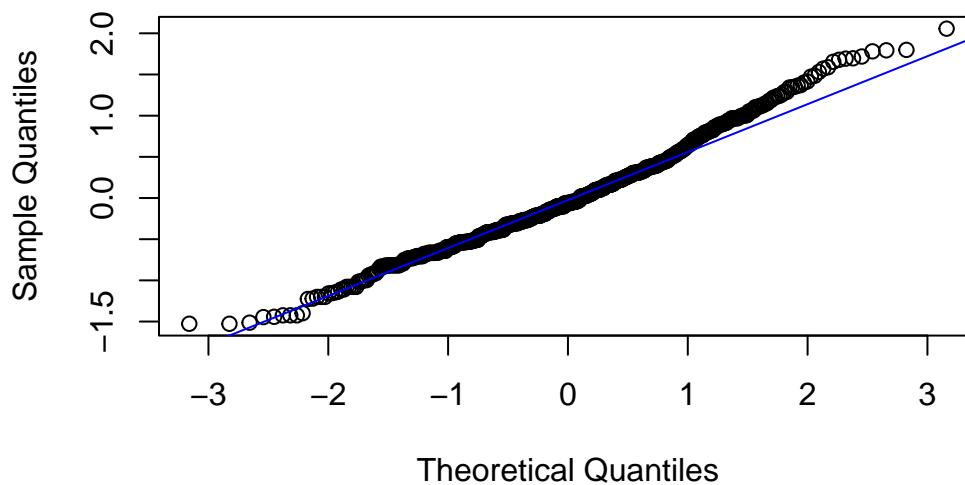
Option 2: Box-Cox Transformation

10

```
# Use lm instead of aov for Box-Cox compatibility
library(MASS)
lm_model <- lm(length_of_stay ~  sex * age_group * cardiohx, data = df_clean)

# Apply Box-Cox transformation
boxcox_result <- boxcox(lm_model, lambda = seq(-2, 2, 0.1),
                        main = "Box-Cox Transformation")
```

Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :
 extra argument 'main' will be disregarded
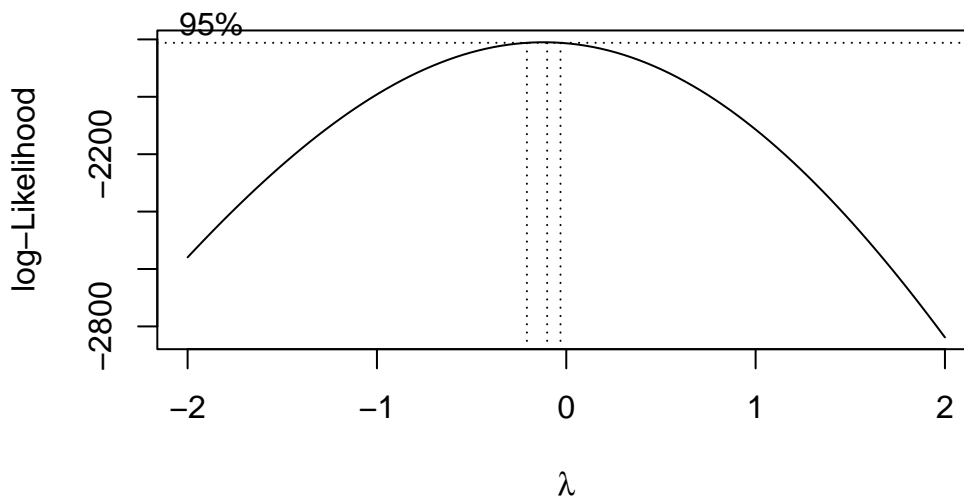

```
# Find optimal lambda
best_lambda <- boxcox_result$x[which.max(boxcox_result$y)]
best_lambda
```

[1] -0.1010101

Since `lambda` ≈ 0, then **log transformation** is best.

1. Three-Way ANOVA: Effects of Sex, Age Group, and Cardiac History on Hospital Length of Stay (Full Model)

```
# Log-transform length_of_stay (add 1 to avoid log(0))
df_clean$log_los <- log(df_clean$length_of_stay + 1)

# Fit the 3-way ANOVA model using the log-transformed data
anova_full <- aov(log_los ~ sex * age_group * cardiohx, data = df_clean)

# View the summary
summary(anova_full)
```

```
                        Df Sum Sq Mean Sq F value Pr(>F)
sex                      1   0.10  0.1044   0.254 0.6143
age_group                3   1.21  0.4028   0.981 0.4013
cardiohx                 1   1.01  1.0135   2.468 0.1167
sex:age_group            3   0.16  0.0522   0.127 0.9440
sex:cardiohx             1   0.23  0.2288   0.557 0.4557
age_group:cardiohx       3   4.06  1.3547   3.299 0.0201 *
sex:age_group:cardiohx   3   0.72  0.2387   0.581 0.6275
Residuals              618 253.80  0.4107
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Key Findings:

- Most **main effects and two-way interactions** were **not statistically significant** ($p > 0.05$).

- However, the **interaction between age group and cardiac history** was **statistically significant** ($F = 3.299$, $p = 0.0201$), indicating that the impact of cardiac history on length of stay **varies by age group**.

- The **three-way interaction** was not significant ($p = 0.6275$), suggesting no combined effect of sex, age, and cardiac history on LOS.

2. Two-Way ANOVA for Age Group $\times$ Cardiac History (Reduced Model)

Since the **three-way ANOVA** showed that only the **interaction between age group and cardiac history** is significant, it's appropriate to fit a **reduced Two-Way ANOVA model** including just those two variables and their interaction.

```
# Fit reduced two-way ANOVA model
anova_reduced <- aov(log_los ~ age_group * cardiohx, data = df_clean)

# Summary of the model
summary(anova_reduced)
```

```
                  Df Sum Sq Mean Sq F value Pr(>F)
age_group          3   1.12  0.3739   0.916 0.4328
cardiohx           1   1.15  1.1516   2.821 0.0936 .
age_group:cardiohx 3   3.44  1.1471   2.810 0.0388 *
Residuals        626 255.58  0.4083
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Key Interpretation:

- **No main effect** of age group or cardiac history alone on log-transformed length of stay.

- The **interaction between age group and cardiac history is significant** ($p = 0.0388$), meaning: The effect of cardiac history on length of stay **differs by age group**, or vice versa.

ANOVA model comparison

```
anova(anova_reduced, anova_full)
```

```
Analysis of Variance Table

Model 1: log_los ~ age_group * cardiohx
Model 2: log_los ~ sex * age_group * cardiohx
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1    626 255.58
2    618 253.80  8    1.7773 0.541 0.8258
```

This test if the reduced model provides a **significantly better fit** than the full odel. Since the p-value is **not significant** and Residual Sum of Squares (RSS) barely changed, then the **reduced model is sufficient**.

AIC/ BIC Comparison

```
# Compare AIC values
AIC(anova_reduced, anova_full)
```

```
              df      AIC
anova_reduced  9 1241.207
anova_full    17 1252.783
```

```
# Compare BIC values
BIC(anova_reduced, anova_full)
```

```
              df      BIC
anova_reduced  9 1281.276
anova_full    17 1328.468
```

Interpretation of AIC/ BIC comparisons:

The reduced model (which includes only age group, cardiac history, and their interaction) has a lower AIC than the full model by **11.57 points**. A difference in AIC greater than 10 is considered **strong evidence** in favor of the simpler model. Based on this AIC comparison, the reduced model is **strongly preferred** over the full model. Removing the non-significant terms involving sex improves **model parsimony** without sacrificing model performance.

Similarly, the **Bayesian Information Criterion (BIC)** further supports this conclusion. The reduced model has a BIC of **1281.28**, while the full model's BIC is **1328.47**, yielding a difference of over **47 points**. Since BIC penalizes model complexity more heavily than AIC, this substantial gap provides **very strong evidence** that the reduced model offers a more efficient and generalizable fit.

Model Comparison Table

Model	Residual DF	Residual SS	AIC	BIC	ANOVA p-value
Full Model	618	253.80	1252.78	1328.47	0.8258
Reduced Model	626	255.58	1241.21	1281.28	—

Model Decision:

The **reduced model** (including only age group, cardiac history, and their interaction) is preferred. It has **lower AIC and BIC values**, and the **ANOVA comparison** shows no significant loss of fit when the non-significant sex-related terms are removed ($p = 0.8258$).

Tukey HSD Post-Hoc Analysis of Age Groups on Hospital Length of Stay

based on reduced model:

```
# Fit one-way ANOVA on log-transformed outcome by age_group
anova_log <- aov(log_los ~ age_group, data = df_clean)

# Tukey HSD post-hoc test
TukeyHSD(anova_log)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = log_los ~ age_group, data = df_clean)

$age_group
                    diff         lwr        upr       p adj
50-65-<50    -0.08350586 -0.2585489 0.09153719 0.6087007
65-80-<50    -0.06099560 -0.2289056 0.10691444 0.7856357
80+-<50      -0.13245221 -0.3649109 0.10000649 0.4577271
65-80-50-65   0.02251026 -0.1479293 0.19294983 0.9864422
80+-50-65    -0.04894635 -0.2832387 0.18534602 0.9496878
80+-65-80    -0.07145661 -0.3004689 0.15755568 0.8526800
```

Conclusion:

There is **no evidence** that hospital length of stay differs significantly by age group based on Tukey's multiple comparisons. All group differences were small and not statistically meaningful.

We run **separate ANOVAs within levels** of `cardiohx` or `age_group` and then do Tukey on `age_group` inside each subgroup.

Tukey HSD Post-Hoc Analysis: Age Group Differences in Length of Stay Among Patients With Cardiac History

based on reduced model:

```
# --- Subset: Patients WITH cardiac history (cardiohx == "1")
with_cardio_1 <- subset(df_clean, cardiohx == "1")
anova_sub_1_log <- aov(log_los ~ age_group, data = with_cardio_1)
TukeyHSD(anova_sub_1_log)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = log_los ~ age_group, data = with_cardio_1)

$age_group
                    diff        lwr        upr       p adj
50-65-<50     -0.02513487 -0.4720526 0.42178286 0.9988855
65-80-<50     -0.27089314 -0.7187567 0.17697045 0.3984398
80+-<50       -0.28535352 -0.8094110 0.23870397 0.4926420
65-80-50-65   -0.24575827 -0.5631131 0.07159653 0.1883910
80+-50-65     -0.26021865 -0.6782719 0.15783461 0.3724913
80+-65-80     -0.01446038 -0.4335247 0.40460389 0.9997413
```

```
# --- Subset: Patients WITHOUT cardiac history (cardiohx == "0")
with_cardio_0 <- subset(df_clean, cardiohx == "0")
anova_sub_0_log <- aov(log_los ~ age_group, data = with_cardio_0)
TukeyHSD(anova_sub_0_log)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = log_los ~ age_group, data = with_cardio_0)

$age_group
                    diff        lwr        upr       p adj
50-65-<50     -0.11681488 -0.31465163 0.08102187 0.4248586
65-80-<50      0.02146110 -0.16281054 0.20573275 0.9905841
80+-<50       -0.05523578 -0.32592298 0.21545142 0.9527487
65-80-50-65    0.13827598 -0.06419746 0.34074943 0.2937004
80+-50-65      0.06157910 -0.22181288 0.34497109 0.9437186
80+-65-80     -0.07669688 -0.35079117 0.19739741 0.8885284
```

Conclusion:

In both the cardiac history and non-cardiac history subgroups, **Tukey HSD post-hoc comparisons revealed no statistically significant differences** in hospital length of stay across age groups. Overall, **age group alone did not significantly influence hospital stay duration within either cardiac history subgroup**.

Interaction Plot for Age Group × Cardiac History → Length of Stay

```
library(dplyr)
library(ggplot2)

# Summarize log-transformed mean length of stay by age group and cardiac history
plot_data <- df_clean %>%
  group_by(age_group, cardiohx) %>%
  summarise(mean_log_los = mean(log_los, na.rm = TRUE), .groups = "drop")

# Plot interaction
ggplot(plot_data, aes(x = age_group, y = mean_log_los, color = cardiohx, group = cardiohx))
  geom_point(size = 3) +
  geom_line(linewidth = 1) +
  labs(
    title = "Interaction Plot: Age Group × Cardiac History → Log Length of Stay",
    x = "Age Group",
    y = "Mean Log Length of Stay",
    color = "Cardiac History"
  ) +
  theme_minimal()
```

Interaction Plot: Age Group × Cardiac History –> Log Length of
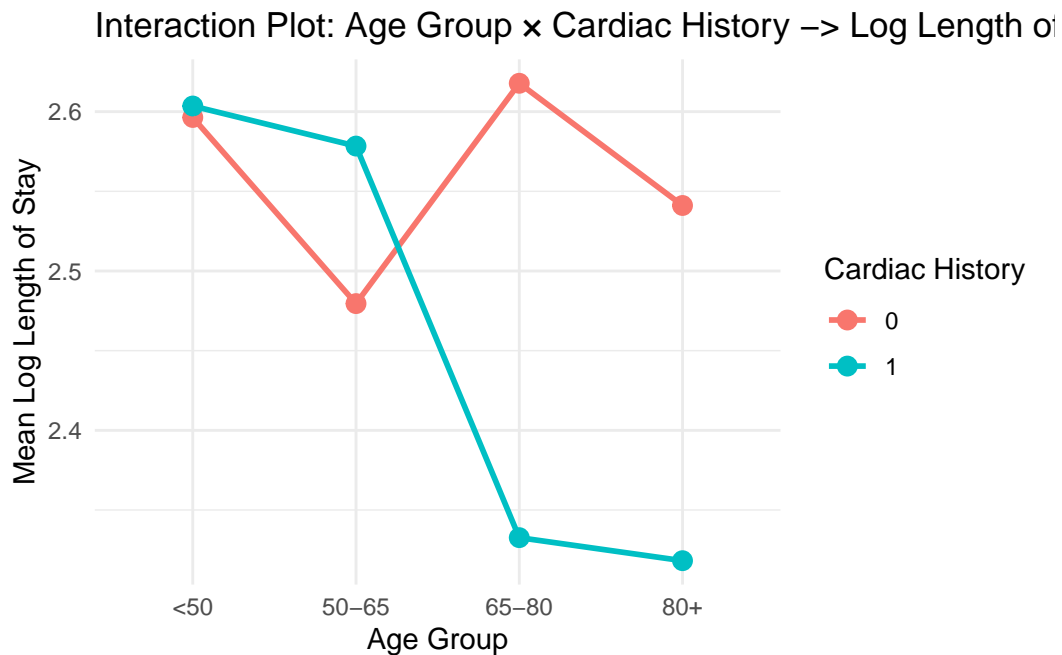
Interpretation of Interaction Plot:

The plot illustrates a **statistically significant interaction** between **age group** and **cardiac history** ($p = 0.0388$), indicating that the effect of age on hospital length of stay depends on whether the patient has a history of cardiovascular disease.

- For patients **without cardiac history** (red line), length of stay is relatively stable but peaks in the **65–80** age group.

- For patients **with cardiac history** (blue line), length of stay **decreases sharply with age**, especially after 50–65.

- This crossover pattern supports a meaningful interaction: the direction and strength of the age effect **differs by cardiac history group**.

Boxplot for Age Group × Cardiac History

```r
library(ggplot2)

ggplot(df_clean, aes(x = age_group, y = log_los, fill = age_group)) +
  geom_boxplot(outlier.shape = 21, outlier.size = 1.5, alpha = 0.8) +
  facet_wrap(~ cardiohx, labeller = label_both) +
```

```
labs(
  title = "Log-Transformed Length of Stay by Age Group and Cardiac History",
  x = "Age Group",
  y = "Log Length of Stay"
) +
theme_minimal() +
theme(legend.position = "none")
```
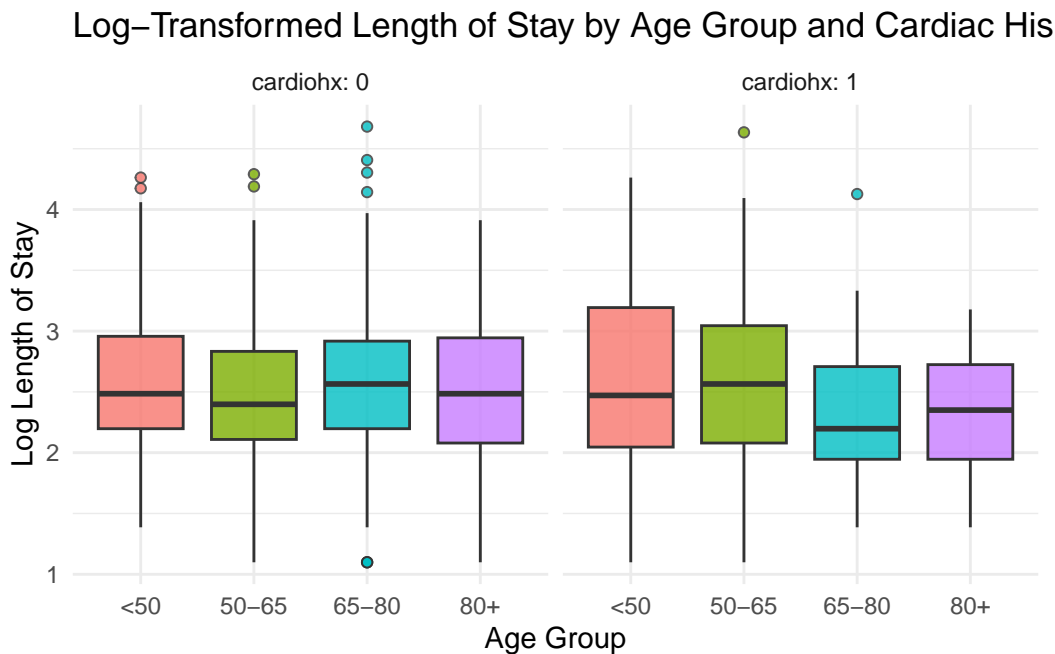

Log Length of Stay by Age Group and Cardiac History:

This plot reveals that patients with a history of cardiovascular disease (right panel) generally show a **steeper decline** in length of stay with increasing age, particularly in the 65–80 group. In contrast, patients **without cardiac history** (left panel) exhibit **more stable median lengths of stay** across age groups, with slightly elevated values in the <50 and 65–80 categories. The variability (IQR and outliers) appears wider in some younger groups, especially for those without cardiac history.

Final Conclusion

This study utilized the Right Heart Catheterization (RHC) dataset from the SUPPORT study to investigate how age group and cardiac history influence hospital length of stay among critically ill patients. After verifying ANOVA assumptions and applying a log transformation to correct for skewness, a three-way ANOVA model was initially fitted including sex, age

group, and cardiac history. However, model comparison via ANOVA and information criteria (AIC and BIC) revealed that a reduced two-way ANOVA model — including only age group, cardiac history, and their interaction — offered equivalent or better model fit with greater parsimony.

The reduced model identified a statistically significant interaction between age group and cardiac history (p = 0.0388), indicating that the effect of age on length of stay depends on cardiac history status. Specifically, length of stay decreased with age among patients with cardiac history, while those without cardiac history showed relatively stable or increasing patterns in older age groups. These trends were confirmed through interaction plots and supported by Tukey HSD post-hoc comparisons, although no pairwise differences were statistically significant.

Visualizations, including boxplots and interaction plots, further illustrated this interaction effect. While sex did not contribute significantly to explaining variation in length of stay, the age group × cardiac history interaction emerged as a meaningful and interpretable driver of variation in hospital resource use.

In summary, this study concludes that the interaction between age group and cardiac history significantly affects hospital length of stay. A reduced two-way ANOVA model provides a more efficient and statistically supported explanation than a full model including sex. These findings highlight the importance of considering both age and comorbid conditions when analyzing healthcare utilization in critical care populations.