

BÁO CÁO ĐỒ ÁN KẾT THÚC MÔN HỌC
MÔ HÌNH HÓA THỐNG KÊ

Lê Nhựt Nam^{1,2} Phạm Thùa Tiêu Thành^{1,2}

¹Khoa Toán - Tin học, Trường Đại học Khoa học Tự nhiên

²Đại học Quốc gia TP.HCM

Ngày 10 tháng 8 năm 2024

MỤC LỤC

LỜI CẢM ƠN	3
DANH MỤC CÁC BẢNG	4
DANH MỤC CÁC HÌNH VẼ, ĐỒ THỊ	5
1 GIỚI THIỆU TỔNG QUAN ĐỒ ÁN	13
1.1 Giới thiệu đồ án	13
1.2 Phân công và kế hoạch thực hiện đồ án	13
1.2.1 Phân công Hoạt động 1 - Dữ liệu Islander	13
1.2.2 Phân công Hoạt động 1 - Dữ liệu CSM	13
1.2.3 Phân công Hoạt động 2 - Dữ liệu về chất lượng rượu	14
1.2.4 Phân công Hoạt động 2 - Dữ liệu về chất lượng lượng không khí	15
1.2.5 Phân công Hoạt động 2 - Dữ liệu về tương tác trên mạng xã hội	16
2 HOẠT ĐỘNG 1	17
2.1 Phân tích tác dụng các loại thuốc chống trầm cảm	17
2.1.1 Giới thiệu chung	17
2.1.2 Phát biểu bài toán	17
2.1.3 Giới thiệu về tập dữ liệu	17
2.1.4 Đọc và phân tích dữ liệu	18
2.1.5 Xử lý dữ liệu	23
2.1.6 Kiểm định các giả thiết thống kê (ANOVA assumptions)	30
2.1.7 Phân tích phương sai k nhân tố	34
2.1.8 Xây dựng và kiểm định mô hình cộng (Additive model)	51
2.1.9 Cải tiến mô hình	54
2.2 Phân tích phim truyền thông và xã hội	82
2.2.1 Giới thiệu chung	82
2.2.2 Phát biểu bài toán	82
2.2.3 Giới thiệu về dữ liệu	82
2.2.4 Khám phá và tiền xử lý dữ liệu	83
2.2.5 Quay lại bước khám phá và tiền xử lý dữ liệu	88
2.2.6 Mô hình hóa	124
2.2.7 Mô hình hóa bằng PCR	139
2.2.8 Mô hình hóa bằng PLS	145

2.2.9	So sánh hai mô hình PCR và PLS	150
3	HOẠT ĐỘNG 2	153
3.1	Phân tích chất lượng rượu	153
3.1.1	Giới thiệu chung	153
3.1.2	Phát biểu bài toán	153
3.1.3	Phân tích chất lượng rượu trắng	153
3.1.4	Phân tích chất lượng rượu đỏ	171
3.1.5	Phân tích chất lượng rượu (bao gồm nhiều tố màu sắc)	190
3.1.6	Kết luận	222
3.2	Phân tích chất lượng không khí	222
3.2.1	Giới thiệu chung	222
3.2.2	Phát biểu bài toán	222
3.2.3	Giới thiệu về dữ liệu	222
3.2.4	Khám phá và tiền xử lý dữ liệu	223
3.2.5	Phân tích đơn biến	225
3.2.6	Phân tích đa biến	277
3.2.7	Mô hình hóa bằng hồi quy tuyến tính đa biến	279
3.2.8	Mô hình hóa bằng PCR	290
3.2.9	Mô hình hóa bằng PLS	294
3.2.10	So sánh và đánh giá PCR và PLS	298
3.2.11	Cải tiến: Random Forest	298
3.2.12	Cải tiến: Support Vector Machine	302
3.2.13	Kết luận	305
3.3	Phân tích hiệu quả tương tác của bài đăng trên mạng xã hội facebook	306
3.3.1	Giới thiệu chung	306
3.3.2	Phát biểu bài toán	306
3.3.3	Giới thiệu về tập dữ liệu	307
3.3.4	Đọc và phân tích dữ liệu	308
3.3.5	Kiểm định các giả thiết thống kê (ANOVA assumptions)	316
3.3.6	Phân tích phương sai k nhân tố	321
3.3.7	Xây dựng và kiểm định mô hình cộng (Additive model)	337
3.3.8	Cải tiến mô hình	341
4	TỔNG KẾT ĐỒ ÁN	367
4.1	Kết luận	367
TÀI LIỆU THAM KHẢO		368

LỜI CẢM ƠN

Lời đầu tiên, chúng tôi xin phép gửi lời cảm ơn chân thành đến giáo viên hướng dẫn của môn học Mô hình hóa thống kê - TS. Nguyễn Thị Mộng Ngọc - giảng viên Bộ môn Xác suất thống kê, Khoa Toán - Tin học, Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia TP. HCM đã trực tiếp hướng dẫn và giúp đỡ tận tình trong suốt quá trình nghiên cứu thực hiện tiểu luận này. Nhờ vào những định hướng, và những kiến thức quý giá được truyền tải của cô, chúng tôi đã hoàn thành trọng vẹn đề tài tiểu luận của mình.

Tiếp theo, chúng tôi xin gửi lời cảm ơn đến quý Thầy, Cô trong khoa Toán - Tin học, trường Đại học Khoa học Tự nhiên, Đại học Quốc gia TP. HCM đã nhiệt tình giảng dạy đã truyền đạt cho tôi những kiến thức sâu sắc về mặt chuyên môn lý thuyết và ứng dụng thực tiễn trong suốt quá trình học tập ở trường. Những điều này đã góp phần quan trọng trong việc hoàn thành tiểu luận này của chúng tôi.

Lê Nhựt Nam Phạm Thùa Tiểu Thành

DANH MỤC CÁC BẢNG

Bảng 1.1	Bảng phân công Hoạt động 1 - Dữ liệu Islander	13
Bảng 1.2	Bảng phân công Hoạt động 1 - Dữ liệu CSM	14
Bảng 1.3	Bảng phân công Hoạt động 2 - Dữ liệu về chất lượng rượu (rượu vang trắng) . .	14
Bảng 1.4	Bảng phân công Hoạt động 2 - Dữ liệu về chất lượng rượu (rượu vang đỏ) . . .	15
Bảng 1.5	Bảng phân công Hoạt động 2 - Dữ liệu về chất lượng rượu (dữ liệu tổng hợp cả trắng và đỏ)	15
Bảng 1.6	Bảng phân công Hoạt động 2 - Dữ liệu về chất lượng không khí	16
Bảng 1.7	Bảng phân công Hoạt động 2 - Dữ liệu về tương tác mạng xã hội	16
Bảng 1.8	Kết hoạch thực hiện nghiên cứu đồ án.	16
Bảng 3.1	Ý nghĩa các cột của dữ liệu chất lượng không khí.	223
Bảng 3.2	Danh sách các biến và thông tin tương ứng	308

DANH MỤC CÁC HÌNH VẼ, ĐỒ THỊ

Hình 2.1	Visualize các biến ‘Age’ và ‘Diff’	21
Hình 2.2	Đồ thị phân phối chuẩn của biến ‘Diff’	22
Hình 2.3	Đồ thị phân phối chuẩn của biến ‘Age’	22
Hình 2.4	Phân phối dữ liệu của từng nhóm Dosage	26
Hình 2.5	Phân phối dữ liệu của từng nhóm Drug	26
Hình 2.6	Đồ thị phân phối chuẩn của biến ‘Diff’	28
Hình 2.7	Biểu đồ phần dư	31
Hình 2.8	Kiểm định độc lập phần dư	33
Hình 2.9	Kết quả ảnh hưởng đơn giữa Drug và Dosage	36
Hình 2.10	Ảnh hưởng đơn giữa Drug và Dosage.	37
Hình 2.11	Tương tác giữa nhóm A và S ở mỗi liều lượng	38
Hình 2.12	Tương tác giữa nhóm A và T ở mỗi liều lượng	38
Hình 2.13	Tương tác giữa nhóm S và T ở mỗi liều lượng	39
Hình 2.14	Tương tác giữa nhóm thấp và nhóm trung bình	40
Hình 2.15	Tương tác giữa nhóm thấp và cao	40
Hình 2.16	Tương tác giữa nhóm trung bình và cao	41
Hình 2.17	Phân tích t-test các nhóm	46
Hình 2.18	Đồ thị phần dư của ảnh hưởng chính giữa Drug và Diff	48
Hình 2.19	Phân tích trung bình giữa các nhóm	50
Hình 2.20	Kết quả mô tả của mô hình tuyến tính	51
Hình 2.21	Trực quan hóa dữ liệu của biến Diff	57
Hình 2.22	Phân phối của phần dư sau khi cài tiến	59
Hình 2.23	Đồ thị Residuals	61
Hình 2.24	Tương tác giữa Drug và Dosage	63
Hình 2.25	Kết quả ảnh hưởng đơn của liều lượng ở mỗi loại thuốc	64
Hình 2.26	Kết quả ảnh hưởng đơn của thuốc ở mỗi liều lượng	65
Hình 2.27	Ảnh hưởng giữa nhóm A và S	66
Hình 2.28	Ảnh hưởng giữa nhóm A và T	66
Hình 2.29	Ảnh hưởng giữa S và T	67

Hình 2.30	Ảnh hưởng giữa nhóm thấp và trung bình	68
Hình 2.31	Ảnh hưởng giữa nhóm thấp và cao	68
Hình 2.32	Tương tác giữa nhóm trung bình vào cao	69
Hình 2.33	Shapiro-test	71
Hình 2.34	Kiểm định độc lập phần dư	73
Hình 2.35	Kiểm định trung bình	74
Hình 2.36	Kết quả t-test	76
Hình 2.37	Shapiro-Wilk test và đồ thị phần dư	78
Hình 2.38	Kiểm định tính độc lập của phần dư	79
Hình 2.39	Tỷ lệ các thể loại phim.	85
Hình 2.40	Phân phối trước và sau khi điền các giá trị bị thiếu đối với các cột Budget, Screens và AggregateFollowers.	87
Hình 2.41	Phân phối ban đầu của Ratings.	88
Hình 2.42	Phân phối sau khi log-scale của Ratings.	89
Hình 2.43	Log-likelihood với các giá trị λ của Ratings.	90
Hình 2.44	Phân phối trước và sau khi biến đổi của Gross.	91
Hình 2.45	Phân phối ban đầu của Budget.	92
Hình 2.46	Phân phối sau khi log-scale của Budget.	93
Hình 2.47	Log-likelihood với các giá trị λ của Budget.	94
Hình 2.48	Phân phối trước và sau khi biến đổi của Budget.	95
Hình 2.49	Phân phối ban đầu của Screens.	96
Hình 2.50	Phân phối sau khi log-scale của Screens.	97
Hình 2.51	Log-likelihood với các giá trị λ của Screens.	98
Hình 2.52	Phân phối trước và sau khi biến đổi của Screens.	99
Hình 2.53	Phân phối ban đầu của Sequel.	100
Hình 2.54	Phân phối ban đầu của Sentiment.	101
Hình 2.55	Phân phối sau khi log-scale của Sentiment.	102
Hình 2.56	Log-likelihood với các giá trị λ của Sentiment.	103
Hình 2.57	Phân phối trước và sau khi biến đổi của Sentiment.	104
Hình 2.58	Phân phối ban đầu của Views.	105
Hình 2.59	Phân phối sau khi log-scale của Views.	106
Hình 2.60	Log-likelihood với các giá trị λ của Views.	107
Hình 2.61	Phân phối trước và sau khi biến đổi của Views.	108
Hình 2.62	Phân phối ban đầu của Likes.	109

Hình 2.63 Phân phối sau khi log-scale của Likes.	110
Hình 2.64 Log-likelihood với các giá trị λ của Likes.	111
Hình 2.65 Phân phối trước và sau khi biến đổi của Likes.	112
Hình 2.66 Phân phối ban đầu của Likes.	113
Hình 2.67 Phân phối sau khi log-scale của Comments.	114
Hình 2.68 Log-likelihood với các giá trị λ của Comments.	115
Hình 2.69 Phân phối trước và sau khi biến đổi của Comments.	116
Hình 2.70 Phân phối ban đầu của Gross.	117
Hình 2.71 Phân phối sau khi log-scale của Gross.	118
Hình 2.72 Log-likelihood với các giá trị λ của Gross.	119
Hình 2.73 Phân phối trước và sau khi biến đổi của Gross.	120
Hình 2.74 Phân phối ban đầu của Gross.	121
Hình 2.75 Phân phối sau khi logscale của AggregateFollowers.	122
Hình 2.76 Log-likelihood với các giá trị λ của AggregateFollowers.	123
Hình 2.77 Phân phối trước và sau khi biến đổi của AggregateFollowers.	124
Hình 2.78 Ma trận tương quan giữa các biến trong tập dữ liệu phim truyền thông.	125
Hình 2.79	131
Hình 2.80 Biểu đồ Residuals vs Fitted Plot.	132
Hình 2.81 Normal Q-Q (quantile-quantile) Plot.	133
Hình 2.82 Histogram biến thặng dư của mô hình hồi quy CSM.	134
Hình 2.83 Scale-Location Plot.	135
Hình 2.84 Residuals vs Leverage Plot.	136
Hình 2.85 Cook Distance Plot của mô hình hồi quy CSM.	137
Hình 2.86 Trực quan kết quả dự đoán của mô hình tốt nhất. RMSE = 196.83.	139
Hình 2.87 Giá trị RMSEP với số lượng thành phần chính khác nhau.	141
Hình 2.88 Kết quả dự đoán của mô hình PCR.	142
Hình 2.89 Kiểm định tính chuẩn thặng dư của mô hình PCR.	144
Hình 2.90 Kiểm định đồng nhất phương sai của mô hình PCR.	145
Hình 2.91 Giá trị RMSEP với số lượng thành phần chính khác nhau.	147
Hình 2.92 Kết quả dự đoán của mô hình PLS.	148
Hình 2.93 Kiểm định tính chuẩn thặng dư của mô hình PCR.	149
Hình 2.94 Kiểm định đồng nhất phương sai của mô hình PCR.	150
Hình 3.1 Chất lượng rượu trắng.	155

Hình 3.2 Histogram tính chua (acidity) trong rượu trắng.	156
Hình 3.3 Boxplot tính chua (acidity) trong rượu trắng.	157
Hình 3.4 Phân phối SO ₂ tự do và tổng lượng SO ₂ trong rượu.	158
Hình 3.5 Phân phối tỷ lệ SO ₂ tự do và tổng lượng SO ₂	159
Hình 3.6 Phân phối Lượng muối sunphat trong rượu.	160
Hình 3.7 Phân phối lượng đường còn lại sau khi lên men trong rượu.	161
Hình 3.8 Phân phối phần trăm cồn trong rượu.	162
Hình 3.9 Phân phối mật độ rượu.	163
Hình 3.10 Phân phối lượng muối rượu.	164
Hình 3.11 Biểu đồ tương quan giữa các biến trong tập dữ liệu rượu trắng.	165
Hình 3.12 Histogram của biến thặng dư mô hình.	169
Hình 3.13 Biểu đồ Heteroscedasticity.	170
Hình 3.14 Chất lượng rượu đỏ.	173
Hình 3.15 Histogram tính chua (acidity) trong rượu đỏ.	174
Hình 3.16 Boxplot tính chua (acidity) trong rượu đỏ.	175
Hình 3.17 Phân phối SO ₂ tự do và tổng lượng SO ₂ trong rượu.	176
Hình 3.18 Phân phối tỷ lệ SO ₂ tự do và tổng lượng SO ₂	177
Hình 3.19 Phân phối Lượng muối sunphat trong rượu.	178
Hình 3.20 Phân phối lượng đường còn lại sau khi lên men trong rượu.	179
Hình 3.21 Phân phối phần trăm cồn trong rượu.	180
Hình 3.22 Phân phối mật độ rượu.	181
Hình 3.23 Phân phối lượng muối rượu.	182
Hình 3.24 Biểu đồ tương quan giữa các biến trong tập dữ liệu rượu đỏ.	183
Hình 3.25 Histogram của biến thặng dư mô hình.	187
Hình 3.26 Biểu đồ Heteroscedasticity.	188
Hình 3.27 Kết quả dự đoán trên bộ dữ liệu chất lượng rượu đỏ.	190
Hình 3.28 Chất lượng rượu.	191
Hình 3.29 Histogram tính chua (acidity) trong rượu.	192
Hình 3.30 Boxplot tính chua (acidity) trong rượu.	193
Hình 3.31 Phân phối SO ₂ tự do và tổng lượng SO ₂ trong rượu.	194
Hình 3.32 Phân phối tỷ lệ SO ₂ tự do và tổng lượng SO ₂	195
Hình 3.33 Phân phối Lượng muối sunphat trong rượu.	196
Hình 3.34 Phân phối lượng đường còn lại sau khi lên men trong rượu.	197
Hình 3.35 Phân phối phần trăm cồn trong rượu.	198

Hình 3.36 Phân phối mật độ rượu.	199
Hình 3.37 Phân phối lượng muối rượu.	200
Hình 3.38 Ma trận tương quan giữa các biến trong tập dữ liệu về rượu.	201
Hình 3.39 Mối quan hệ giữa Density và Quality	202
Hình 3.40 Biểu đồ boxplot về mối quan hệ giữa Density và Quality	203
Hình 3.41 Mối quan hệ giữa Alcoholy và Quality	204
Hình 3.42 Mối quan hệ giữa Chlorides và Quality	205
Hình 3.43 Mối quan hệ giữa Volatile Acidity và Quality	206
Hình 3.44 Mối quan hệ giữa tổng lượng SO2 và lượng đường còn lại sau khi lên men	207
Hình 3.45 Mối quan hệ giữa nồng độ cồn và lượng đường đến mật độ rượu	208
Hình 3.46 Mối quan hệ giữa đường dư và lượng đường đến mật độ rượu	209
Hình 3.47 Mối quan hệ giữa màu sắc và mật độ rượu	210
Hình 3.48 Mối quan hệ giữa màu sắc và lượng đường còn lại sau khi lên men	211
Hình 3.49 Mối quan hệ giữa màu sắc và tổng lượng lưu huỳnh trong rượu	212
Hình 3.50 Mối quan hệ giữa màu sắc và lượng lưu huỳnh tự do trong rượu	213
Hình 3.51 Mối quan hệ giữa màu sắc và tính chua của rượu	214
Hình 3.52 Mối quan hệ giữa mật độ và chất lượng rượu dựa trên màu sắc	215
Hình 3.53 Mối quan hệ giữa nồng độ cồn và chất lượng rượu dựa trên màu sắc	216
Hình 3.54 Mối quan hệ giữa lượng muối và chất lượng rượu dựa trên màu sắc	217
Hình 3.55 Mối quan hệ giữa độ chua và chất lượng rượu dựa trên màu sắc	218
Hình 3.56 Phân phối ban đầu của Carbon monoxide.	225
Hình 3.57 Phân phối sau khi log-scale của Carbon monoxide.	226
Hình 3.58 Log-likelihood với các giá trị λ của Carbon monoxide.	227
Hình 3.59 Phân phối trước và sau khi biến đổi của Carbon monoxide.	228
Hình 3.60 Phân phối ban đầu của Sensor response Carbon monoxide.	229
Hình 3.61 Phân phối sau khi log-scale của Sensor response Carbon monoxide.	230
Hình 3.62 Log-likelihood với các giá trị λ của Sensor response Carbon monoxide.	231
Hình 3.63 Phân phối trước và sau khi biến đổi của Sensor response Carbon monoxide.	232
Hình 3.64 Phân phối ban đầu của Non-methane hydrocarbons.	233
Hình 3.65 Phân phối sau khi log-scale của Non-methane hydrocarbons.	234
Hình 3.66 Log-likelihood với các giá trị λ của Non-methane hydrocarbons.	235
Hình 3.67 Phân phối trước và sau khi biến đổi của Non-methane hydrocarbons	236
Hình 3.68 Phân phối ban đầu của Benzene concentration.	237
Hình 3.69 Phân phối sau khi log-scale của Non-methane hydrocarbons.	238

Hình 3.70 Log-likelihood với các giá trị λ của Non-methane hydrocarbons.	239
Hình 3.71 Phân phối trước và sau khi biến đổi của Non-methane hydrocarbons	240
Hình 3.72 Phân phối ban đầu của Sensor response cho NMHC.	241
Hình 3.73 Phân phối sau khi log-scale của Sensor response cho NMHC.	242
Hình 3.74 Log-likelihood với các giá trị λ của Sensor response cho NMHC.	243
Hình 3.75 Phân phối trước và sau khi biến đổi của Sensor response cho NMHC	244
Hình 3.76 Phân phối ban đầu của Sensor response cho Nitrogen oxides.	245
Hình 3.77 Phân phối sau khi log-scale của Sensor response cho Nitrogen oxides.	246
Hình 3.78 Log-likelihood với các giá trị λ của Sensor response cho Nitrogen oxides.	247
Hình 3.79 Phân phối trước và sau khi biến đổi của Sensor response cho Nitrogen oxides. . .	248
Hình 3.80 Phân phối ban đầu của Sensor response cho Nitrogen oxides.	249
Hình 3.81 Phân phối sau khi log-scale của Sensor response cho Nitrogen oxides.	250
Hình 3.82 Log-likelihood với các giá trị λ của Sensor response cho Nitrogen oxides.	251
Hình 3.83 Phân phối trước và sau khi biến đổi của Sensor response cho Nitrogen oxides. . .	252
Hình 3.84 Phân phối ban đầu của Sensor response cho Nitrogen oxides.	253
Hình 3.85 Phân phối sau khi log-scale của Sensor response cho Nitrogen oxides.	254
Hình 3.86 Log-likelihood với các giá trị λ của Sensor response cho Nitrogen oxides.	255
Hình 3.87 Phân phối trước và sau khi biến đổi của Sensor response cho Nitrogen oxides. . .	256
Hình 3.88 Phân phối ban đầu của Sensor response cho Nitrogen oxides.	257
Hình 3.89 Phân phối sau khi log-scale của Sensor response cho Nitrogen oxides.	258
Hình 3.90 Log-likelihood với các giá trị λ của Sensor response cho Nitrogen oxides.	259
Hình 3.91 Phân phối trước và sau khi biến đổi của Sensor response cho Nitrogen oxides. . .	260
Hình 3.92 Phân phối ban đầu của Sensor response cho Nitrogen oxides.	261
Hình 3.93 Phân phối sau khi log-scale của Sensor response cho Nitrogen oxides.	262
Hình 3.94 Log-likelihood với các giá trị λ của Sensor response cho Nitrogen oxides.	263
Hình 3.95 Phân phối trước và sau khi biến đổi của Sensor response cho Nitrogen oxides. . .	264
Hình 3.96 Phân phối ban đầu của Sensor response cho Nitrogen oxides.	265
Hình 3.97 Phân phối sau khi log-scale của Sensor response cho Nitrogen oxides.	266
Hình 3.98 Log-likelihood với các giá trị λ của Sensor response cho Nitrogen oxides.	267
Hình 3.99 Phân phối trước và sau khi biến đổi của Sensor response cho Nitrogen oxides. . .	268
Hình 3.100 Phân phối ban đầu của Sensor response cho Nitrogen oxides.	269
Hình 3.101 Phân phối sau khi log-scale của Sensor response cho Nitrogen oxides.	270
Hình 3.102 Log-likelihood với các giá trị λ của Sensor response cho Nitrogen oxides.	271
Hình 3.103 Phân phối trước và sau khi biến đổi của Sensor response cho Nitrogen oxides. . .	272

Hình 3.104	Phân phối ban đầu của Sensor response cho Nitrogen oxides.	273
Hình 3.105	Phân phối sau khi log-scale của Sensor response cho Nitrogen oxides.	274
Hình 3.106	Log-likelihood với các giá trị λ của Sensor response cho Nitrogen oxides.	275
Hình 3.107	Phân phối trước và sau khi biến đổi của Sensor response cho Nitrogen oxides.	276
Hình 3.108	Matrận tương quan giữa các biến trong tập dữ liệu Air.	277
Hình 3.109	.	283
Hình 3.110	Biểu đồ Residuals vs Fitted Plot của mô hình hồi quy Air.	284
Hình 3.111	Normal Q–Q (quantile-quantile) Plot cho mô hình hồi quy Air.	285
Hình 3.112	Histogram biến thặng dư của mô hình hồi quy Air.	286
Hình 3.113	Scale-Location Plot cho mô hình hồi quy Air.	287
Hình 3.114	Residuals vs Leverage Plot cho mô hình hồi quy Air.	288
Hình 3.115	Cook Distance Plot của mô hình hồi quy CSM.	289
Hình 3.116	Trực quan kết quả dự đoán của mô hình Air tốt nhất. RMSE = 0.5201.	290
Hình 3.117	Giá trị RMSEP với số lượng thành phần chính khác nhau của mô hình Air PCR.	292
Hình 3.118	Kết quả dự đoán của mô hình Air PCR.	293
Hình 3.119	Giá trị RMSEP với số lượng thành phần chính khác nhau của mô hình Air PLS.	296
Hình 3.120	Kết quả dự đoán của mô hình Air PLS.	297
Hình 3.121	Kết quả dự đoán của mô hình Air Random Forest.	299
Hình 3.122	Mức độ quan trọng của đặc trưng từ mô hình Air Random Forest.	300
Hình 3.123	Giải thích kết quả từ mô hình Air Random Forest.	302
Hình 3.124	Kết quả dự đoán của mô hình Air SVM.	303
Hình 3.125	Giải thích kết quả từ mô hình Air SVM.	305
Hình 3.126	Tóm tắt dữ liệu Facebook.	309
Hình 3.127	Khảo sát outliers.	311
Hình 3.128	Biểu đồ phân phối biến like.	312
Hình 3.129	Biểu đồ phân bố biến Paid.	313
Hình 3.130	Biểu đồ phân bố biến Category.	314
Hình 3.131	Phân phối biến Like.	315
Hình 3.132	Biểu đồ phần dư.	318
Hình 3.133	Kết quả kiểm định durbinWatsonTest.	320
Hình 3.134	Tương tác giữa các biến trong mô hình.	322
Hình 3.135	Kết quả tương tác giữa Paid và Category.	323
Hình 3.136	Ảnh hưởng đơn giản giữa thể loại ở việc thuê quảng cáo.	324
Hình 3.137	Tương tác giữa nhóm 1 và 2 ở mỗi liều lượng.	325

Hình 3.138	Tương tác giữa nhóm 1 và 3	325
Hình 3.139	Tương tác giữa nhóm 2 và 3	326
Hình 3.140	Ảnh hưởng đơn giữa quảng cáo ứng với các nhóm thể loại	326
Hình 3.141	Kết quả kiểm định Shapiro-Wilk test	329
Hình 3.142	Kiểm định durbinWatsonTest	330
Hình 3.143	Kiểm định Tukey's	332
Hình 3.144	Kết quả Shapiro-Wilk test và đồ thị phân phối	334
Hình 3.145	Kiểm định durbinWatsonTest	336
Hình 3.146	Kết quả kiểm định giá trị trung bình	337
Hình 3.147	Shapiro test và biểu đồ chuẩn của phần dư	339
Hình 3.148	Biểu đồ trước khi loại bỏ ngoại lai	345
Hình 3.149	Biểu đồ sau khi loại bỏ ngoại lai	346
Hình 3.150	Shapiro-Wilk normality test và biểu đồ phần dư	348
Hình 3.151	Đồ thị Residuals	349
Hình 3.152	Tương tác giữa Category và Paid	350
Hình 3.153	Shapiro-test	353
Hình 3.154	Kiểm định độc lập phần dư	355
Hình 3.155	Kiểm định trung bình	356
Hình 3.156	Shapiro-Wilk normality test và độ thị phân phối	359
Hình 3.157	Kiểm định tính độc lập của phần dư	360
Hình 3.158	Kiểm định độ hiệu quả trung bình	361
Hình 3.159	Shapiro test và biểu đồ phân phối	364

CHƯƠNG 1

GIỚI THIỆU TỔNG QUAN ĐỒ ÁN

1.1. Giới thiệu đồ án

Báo cáo này là bài viết tổng hợp quá trình thực hiện đồ án môn học Mô hình hóa thống kê. Đồ án này bao gồm hai hoạt động:

- Hoạt động 1: Thực hiện các yêu cầu cho tập dữ liệu Islander và CSM.
- Hoạt động 2: Tự chọn 3 bộ dữ liệu và thực hiện lại các yêu cầu.

1.2. Phân công và kế hoạch thực hiện đồ án

1.2.1. Phân công Hoạt động 1 - Dữ liệu Islander

Bảng 1.1: Bảng phân công Hoạt động 1 - Dữ liệu Islander

STT	Công việc	Người thực hiện	Kết quả
1	Đọc dữ liệu và tiền xử lý	Nam	Hoàn thành
2	Kiểm định các giả thiết thống kê	Thành	Hoàn thành
3	Phân tích phương sai k nhân tố	Thành	Hoàn thành
4	Xây dựng và kiểm định mô hình cộng	Nam	Hoàn thành
5	Cải tiến mô hình	Thành	Hoàn thành

1.2.2. Phân công Hoạt động 1 - Dữ liệu CSM

Trong hoạt động này, nhóm đã đan xen xử lý nếu giả thiết mô hình không thỏa thì nhóm đã sử dụng luôn box-cox transformation, loại bỏ ngoại lai bằng Cook Distance.

Bảng 1.2: Bảng phân công Hoạt động 1 - Dữ liệu CSM

STT	Công việc	Người thực hiện	Kết quả
1	Đọc dữ liệu và tiền xử lý	Nam	Hoàn thành
2	Khám phá và tiền xử lý dữ liệu	Thành	Hoàn thành
3	Xử lý missing values (loại bỏ hoàn toàn)	Nam	Hoàn thành
4	Xử lý missing values (điền mean, median, zeros)	Thành	Hoàn thành
5	Xử lý missing values (điền bằng PCA)	Thành+Nam	Hoàn thành
6	Phân tích đơn biến	Thành	Hoàn thành
7	Phân tích đa biến, Khảo sát ngoại lai	Nam	Hoàn thành
8	Mô hình hóa hồi quy tuyến tính đa biến và kiểm định mô hình	Thành	Hoàn thành
9	Mô hình hóa bằng PCR	Thành	Hoàn thành
10	Mô hình hóa bằng PLS	Thành+Nam	Hoàn thành
11	So sánh và đánh giá	Nam	Hoàn thành
12	Dự đoán và trực quan hóa kết quả	Nam	Hoàn thành

1.2.3. Phân công Hoạt động 2 - Dữ liệu về chất lượng rượu

Tập dữ liệu về rượu có 2 tập dữ liệu con: rượu vang trắng, và rượu vang đỏ. Nhóm thực hiện khảo sát từng tập dữ liệu con và sau đó kết hợp lại thành 1 bộ dữ liệu với biến bổ sung "color" thể hiện màu sắc của rượu. Bên cạnh đó, nhóm đã đan xen xử lý nếu giả thiết mô hình không thỏa thì nhóm đã sử dụng luôn box-cox transformation, loại bỏ ngoại lai bằng Cook Distance.

Bảng 1.3: Bảng phân công Hoạt động 2 - Dữ liệu về chất lượng rượu (rượu vang trắng)

STT	Công việc	Người thực hiện	Kết quả
1	Đọc dữ liệu và tiền xử lý	Nam	Hoàn thành
2	Khám phá và tiền xử lý dữ liệu	Thành	Hoàn thành
3	Phân tích đơn biến	Thành	Hoàn thành
4	Phân tích đa biến, Khảo sát ngoại lai	Nam	Hoàn thành
5	Mô hình hóa hồi quy tuyến tính đa biến	Thành	Hoàn thành
6	Kiểm định các giả thiết của mô hình	Thành	Hoàn thành
7	Dự đoán và trực quan hóa kết quả	Thành	Hoàn thành

Bảng 1.4: Bảng phân công Hoạt động 2 - Dữ liệu về chất lượng rượu (rượu vang đỏ)

STT	Công việc	Người thực hiện	Kết quả
1	Đọc dữ liệu và tiền xử lý	Nam	Hoàn thành
2	Khám phá và tiền xử lý dữ liệu	Thành	Hoàn thành
3	Phân tích đơn biến	Thành	Hoàn thành
4	Phân tích đa biến, Khảo sát ngoại lai	Nam	Hoàn thành
5	Mô hình hóa hồi quy tuyến tính đa biến	Nam	Hoàn thành
6	Kiểm định các giả thiết của mô hình	Nam	Hoàn thành
7	Dự đoán và trực quan hóa kết quả	Nam	Hoàn thành

Bảng 1.5: Bảng phân công Hoạt động 2 - Dữ liệu về chất lượng rượu (dữ liệu tổng hợp cả trắng và đỏ)

STT	Công việc	Người thực hiện	Kết quả
1	Đọc dữ liệu và tiền xử lý	Nam	Hoàn thành
2	Khám phá và tiền xử lý dữ liệu	Thành	Hoàn thành
3	Phân tích đơn biến	Thành	Hoàn thành
4	Phân tích ảnh hưởng của các biến đối với chất lượng rượu	Thành	Hoàn thành
5	Phân tích dựa trên màu sắc của rượu	Nam	Hoàn thành
6	Phân tích tương quan giữa các biến dựa trên màu sắc	Thành	Hoàn thành
7	Mô hình hóa hồi quy tuyến tính đa biến	Nam	Hoàn thành
8	Kiểm định các giả thiết của mô hình	Nam	Hoàn thành
9	Dự đoán và trực quan hóa kết quả	Nam	Hoàn thành

1.2.4. Phân công Hoạt động 2 - Dữ liệu về chất lượng lượng không khí

Trong hoạt động này, nhóm đã đan xen xử lý nếu giả thiết mô hình không thỏa thì nhóm đã sử dụng luôn box-cox transformation, loại bỏ ngoại lai bằng Cook Distance.

Bảng 1.6: Bảng phân công Hoạt động 2 - Dữ liệu về chất lượng không khí

STT	Công việc	Người thực hiện	Kết quả
1	Đọc dữ liệu và tiền xử lý	Nam	Hoàn thành
2	Khám phá và tiền xử lý dữ liệu	Thành	Hoàn thành
3	Phân tích đơn biến	Nam	Hoàn thành
4	Phân tích đa biến	Thành	Hoàn thành
5	Mô hình hóa và kiểm định các giả thiết của mô hình hồi quy tuyến tính	Nam	Hoàn thành
6	Mô hình hóa và kiểm định các giả thiết của mô hình PCR	Thành	Hoàn thành
7	Mô hình hóa và kiểm định các giả thiết của mô hình PLS	Thành	Hoàn thành
8	Cải tiến (Random Forest và SVM)	Thành+Nam	Hoàn thành
9	So sánh, trực quan hóa kết quả	Thành+Nam	Hoàn thành

1.2.5. Phân công Hoạt động 2 - Dữ liệu về tương tác trên mạng xã hội

Bảng 1.7: Bảng phân công Hoạt động 2 - Dữ liệu về tương tác mạng xã hội

STT	Công việc	Người thực hiện	Kết quả
1	Đọc dữ liệu và tiền xử lý	Nam	Hoàn thành
2	Kiểm định các giả thiết thống kê	Thành	Hoàn thành
3	Phân tích phương sai k nhân tố	Nam	Hoàn thành
4	Xây dựng và kiểm định mô hình cộng	Thành	Hoàn thành
5	Cải tiến mô hình	Thành	Hoàn thành

Công việc	Thời gian
Nhận đồ án và thực hiện Hoạt động 1	10/07/2024 - 17/07/2024
Tìm dữ liệu và thực hiện Hoạt động 2	19/07/2024 - 26/07/2024
Viết báo cáo	27/07/2024 - 02/07/2024
Điều chỉnh và bổ sung	03/08/2024 - 09/08/2024

Bảng 1.8: Kết hoạch thực hiện nghiên cứu đồ án.

CHƯƠNG 2

HOẠT ĐỘNG 1

2.1. Phân tích tác dụng các loại thuốc chống trầm cảm

2.1.1. Giới thiệu chung

Thuốc chống trầm cảm là một trong những phương pháp điều trị phổ biến cho các rối loạn tâm thần như trầm cảm và lo âu. Tuy nhiên, một trong những lo ngại chính khi sử dụng các loại thuốc này là tác động tiềm tàng của chúng lên chức năng nhận thức, đặc biệt là trí nhớ. Việc nghiên cứu và hiểu rõ ảnh hưởng của thuốc chống trầm cảm đối với trí nhớ không chỉ quan trọng đối với việc tối ưu hóa liệu pháp điều trị mà còn giúp giảm thiểu các tác dụng phụ không mong muốn, cải thiện chất lượng cuộc sống của bệnh nhân.

2.1.2. Phát biểu bài toán

Các loại thuốc benzodiazepin đã cho thấy có tác dụng phá vỡ tác động tích cực của tiềm năng lâu dài giữa các tế bào đối với việc thu hồi trí nhớ và các mối liên hệ đã biết. Bằng cách phân biệt các tác dụng phụ lâu dài của Alprazolam (dài hạn) và Triazolam (ngắn hạn), bệnh nhân có thể được chẩn đoán tốt hơn để giảm thiểu bất kỳ tổn thương nào đối với khả năng siêu nhận thức (metacognition) và thu hồi trí nhớ của não. Nghiên cứu sâu hơn cũng chỉ ra rằng chỉ cần nhở lại những ký ức cụ thể có liên quan đến cảm xúc mạnh mẽ sẽ khiến những cảm xúc đó được hiện thực hóa ở thời điểm hiện tại và ảnh hưởng đến những suy nghĩ trong tương lai trong một khoảng thời gian ngắn (khoảng 10 phút). Sự hiện diện của cảm xúc vui và cảm xúc buồn được quan tâm và được biết là có ảnh hưởng đáng kể đến việc thu hồi trí nhớ, từ đó đặt ra câu hỏi, những ảnh hưởng nào đến hiệu suất thu hồi trí nhớ của các thuốc benzodiazepin sau khi được bắt đầu bằng ký ức vui hay buồn? Nghiên cứu lâm sàng này sẽ cho thấy liệu tâm trạng của ký ức hỗ trợ hoặc cản trở việc nhớ lại trí nhớ có độc lập với các yếu tố khác hay không, nếu hiệu quả của thuốc benzodiazepin không chỉ phụ thuộc vào khả năng chịu đựng của người tham gia mà còn cả tâm trạng của họ, và cuối cùng là khả năng tăng cường hoặc làm giảm hiệu suất nhớ lại trí nhớ khi được kết hợp cùng nhau vượt ra ngoài phản ứng đã biết với việc sử dụng thuốc benzodiazepin hoặc ký ức liên quan đến tâm trạng của riêng bệnh nhân.

2.1.3. Giới thiệu về tập dữ liệu

Dữ liệu được cho trong tập tin “Islander-data.csv” lấy từ <https://www.kaggle.com/datasets/steveahn/memory-test-on-drugged-islanders-data>

Dữ liệu chứa thông tin về một thử nghiệm về tác dụng phụ của các loại thuốc chống trầm cảm đối

với trí nhớ của người tham gia thử nghiệm, được đánh giá thông qua thời gian hoàn thành một bài kiểm tra trí nhớ. Người tham gia thử nghiệm sẽ được sử dụng một trong ba loại thuốc khác nhau, với 3 hàm lượng khác nhau và sẽ tiếp xúc với các ký ức vui hoặc buồn trong vòng 10 phút trước khi tiến hành kiểm tra. Thời gian hoàn thành bài kiểm tra của người tham gia sẽ được ghi nhận trước và sau khi kết thúc thử nghiệm để đánh giá hiệu quả của từng loại thuốc cũng như hàm lượng thuốc khác nhau. (Những người này đều trên 25 tuổi nhằm đảm bảo thuỷ trán phát triển hoàn thiện, nơi đảm nhận chức năng nhận thức và gợi lại ký ức). Dữ liệu được thu thập bởi ông Almohalwas tại UCLA bao gồm 198 quan trắc với 9 biến sau:

- **first-name:** tên của người tham gia thử nghiệm
- **last-name:** họ của người tham gia thử nghiệm
- **HappySadgroup:** loại ký ức được tiếp xúc trước khi kiểm tra (H: vui, S: buồn)
- **Dosage:** Mức độ hàm lượng thuốc sử dụng (1: thấp, 2: trung bình, 3: cao)
- **Drug:** Loại thuốc sử dụng (A: , Alprazolam, T: Triazolam, S: Placebo)
- **Mem-Score-Before:** Thời gian (giây) cần để hoàn thành bài kiểm tra trước khi tiếp xúc với thuốc chữa trầm cảm
- **Mem-Score-After:** Thời gian (giây) cần để hoàn thành bài kiểm tra sau khi tiếp xúc với thuốc chữa trầm cảm
- **Diff:** Chênh lệch giữa thời gian (giây) hoàn thành bài kiểm tra trước và sau khi sử dụng thuốc.

2.1.4. Đọc và phân tích dữ liệu

Ở bước này, chúng ta sẽ thực hiện một số công việc chính như sau:

- 1 . Đọc dữ liệu và nhận xét tổng quan
- 2 . Thực hiện kiểm tra về bộ dữ liệu bao gồm: Kiểm tra tính độc lập, Kiểm tra dữ liệu khuyết, và kiểm tra outliers của bộ dữ liệu.
- 4 . Trực quan hóa dữ liệu và rút ra nhận xét.

Ngôn ngữ được sử dụng xuyên suốt trong toàn bộ bài báo cáo là R.

Bước 1 : Đọc dữ liệu và nhận xét tổng quan

```

1 data_path = "/content/Islander_data.csv"
2 islander_raw = read.csv(data_path, header = TRUE, sep = ",",
3                         , stringsAsFactors = FALSE)
4 str(islander_raw)
5 names(islander_raw)
6 dim(islander_raw)

```

Kết quả trả về như sau:

```

1 'data.frame': 198 obs. of 9 variables:
2   $ first_name     : chr  "Bastian" "Evan" "Florencia" "
3   "Holly" ...
4   $ last_name      : chr  "Carrasco" "Carrasco" "Carrasco"
5   "Carrasco" ...
6   $ age            : int  25 52 29 50 52 37 35 38 29 36 ...
7   $ Happy_Sad_group: chr  "H" "S" "H" "S" ...
8   $ Dosage          : int  1 1 1 1 1 1 1 1 1 1 ...
9   $ Drug            : chr  "A" "A" "A" "A" ...
10  $ Mem_Score_Before: num  63.5 41.6 59.7 51.7 47 66.4 44.1
11  76.3 56.2 54.8 ...
12  $ Mem_Score_After : num  61.2 40.7 55.1 51.2 47.1 58.1 56
13  74.8 45 75.9 ...
14  $ Diff            : num  -2.3 -0.9 -4.6 -0.5 0.1 -8.3 11.9
15  -1.5 -11.2 21.1 ...
16  'first_name' 'last_name' 'age' 'Happy_Sad_group' 'Dosage' 'Drug'
17  'Mem_Score_Before' 'Mem_Score_After' 'Diff'
18  1989

```

Nhìn vào từng biến hiện thị, ta có một số nhận xét như sau:

- Các biến **first-name** và **last-name** chứa thông tin về tên của người khảo sát (kiểu dữ liệu character), về mặt thống kê biến này không có ý nghĩa nên sẽ được loại bỏ khỏi dữ liệu khi khảo sát.
- Các biến **HappySadgroup**, **Dosage** và **Drug** được thể hiện dưới dạng category (nhóm) vì thế sẽ được asFactor trước khi khảo sát.
- Các biến **Mem-Score-Before**, **Mem-Score-After**, và **Diff** được thể hiện dưới dạng kiểu dữ liệu numeric, tuy nhiên, ở đây ta có **Diff = Mem-Score-Before - Mem-Score-After** (đa cộng tuyến), vì vậy ta chỉ cần khảo sát biến phụ thuộc **Diff**, các biến còn lại sẽ loại bỏ ra khỏi dữ liệu trước khi khảo sát.

- Biến **age** kiểu dữ liệu int, chứa thông tin về tuổi của người khảo sát, dao động từ 24 tuổi đến 83 tuổi. Thay vì khảo sát trên từng nhóm độ tuổi riêng biệt (rất nhiều), ta sẽ tiến hành chia thành 2 nhóm chính là nhóm tuổi < 50 và nhóm còn lại.

Bước 2 : Thực hiện kiểm tra về bộ dữ liệu bao gồm: Kiểm tra tính độc lập, Kiểm tra dữ liệu khuyết, và kiểm tra outliers của bộ dữ liệu.

- Kiểm tra tính độc lập của dữ liệu

```

1 duplicates = islander_raw[duplicated(islander_raw), ]
2 duplicate_counts = table(islander_raw[duplicated(
    islander_raw), ])
3 print(duplicates)
4 print(duplicate_counts)

```

Thực thi đoạn mã trên, ta thấy rằng đối với bộ dữ liệu này, không có sự trùng lặp giữa các quan trắc, vậy chúng độc lập với nhau.

```

1 < table of extent 0 x 0 x 0 x 0 x 0 x 0 x 0 x 0 x 0 x 0 >

```

- Kiểm tra dữ liệu khuyết

```

1 missing_ratio = function(s) {
2   round(mean(is.na(s)) * 100, 1)
3 }
4 sapply(islander_raw, missing_ratio)

```

Thực thi đoạn mã trên, ta thấy rằng đối với bộ dữ liệu này, các quan trắc không có khuyết đặc trưng ở tất cả các quan trắc.

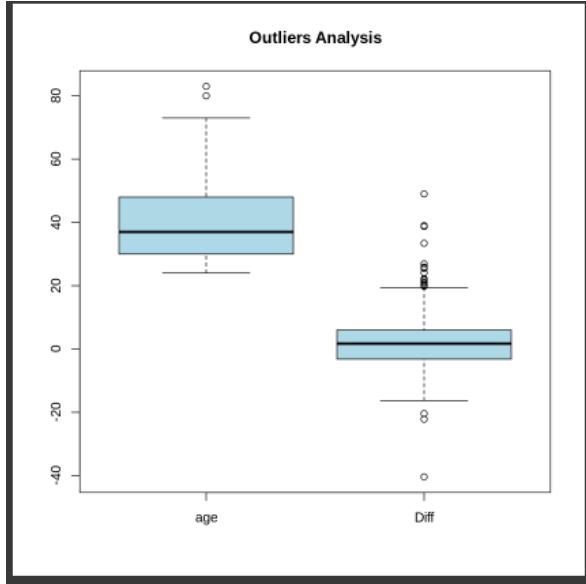
first_name	last_name	age
0	0	0
Happy_Sad_group	Dosage	Drug
0	0	0
Mem_Score_Before	Mem_Score_After	Diff
0	0	0

- Kiểm tra ngoại lai và cực ngoại lai Đối với bước này, ta chỉ kiểm tra đối với các biến có giá trị là numerics, như vậy ta sẽ khảo sát các biến age và Diff

```

1 # Create a box plot
2 boxplot(islander_raw[c("age", "Diff")], main="Outliers
    Analysis", col="lightblue")

```



Hình 2.1: Visualize các biến ‘Age’ và ‘Diff’

Từ biểu đồ hộp, ta có nhận xét sau đây:

- * **age**: Có một vài điểm cực ngoại lai ở phía trên.
- * **Diff**: Tồn tại nhiều điểm cực ngoại lai ở trên và ở phía dưới box.

Bước 4 : Trực quan hóa dữ liệu và rút ra nhận xét.

Ta sẽ dùng R để vẽ ra biểu đồ phân bố chuẩn của dữ liệu

```

1 # Biến Age
2 ggplot(islander_raw, aes(x = age)) +
3   geom_histogram(aes(y = ..density..), bins = 30, color = "black",
4                 fill = "lightblue") +
5   geom_density(alpha = 0.2, fill = "#FF6666") +
6   stat_function(fun = dnorm, args = list(mean = mean(
7     islander_raw$age, na.rm = TRUE), sd = sd(islander_raw$age,
8     na.rm = TRUE)),
9                 color = "blue", size = 1) +
10  theme_minimal() +
11  labs(title = "Histogram of age variable", x = "age", y =
12    "Density")

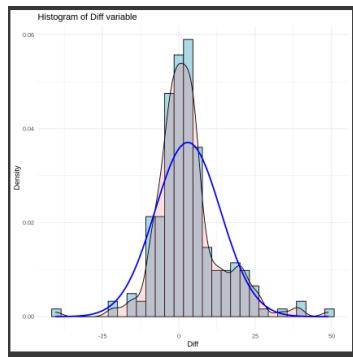
# Biến Diff
ggplot(islander_raw, aes(x = age)) +
  geom_histogram(aes(y = ..density..), bins = 30, color = "black",
                 fill = "lightblue") +

```

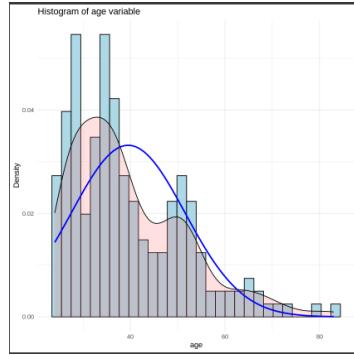
```

13  geom_density(alpha = 0.2, fill = "#FF6666") +
14  stat_function(fun = dnorm, args = list(mean = mean(
15      islander_raw$age, na.rm = TRUE), sd = sd(islander_raw$age,
16      na.rm = TRUE)),
17      color = "blue", size = 1) +
  theme_minimal() +
  labs(title = "Histogram of age variable", x = "age", y =
    "Density")

```



Hình 2.2: Đồ thị phân phối chuẩn của biến ‘Diff’



Hình 2.3: Đồ thị phân phối chuẩn của biến ‘Age’

Từ hai đồ thị trên, ta có một số nhận xét như sau: - ‘Age’: Không có dạng phân bố chuẩn, lệch trái so với giá trị trung bình - ‘Diff’: Có dạng phân bố gần chuẩn.

Bước 4 : Kết luận

Sau khi hoàn thành bước khảo sát dữ liệu, ta có một số kết luận như sau

- Các biến first-name và last-name chứa thông tin về tên của người khảo sát (kiểu dữ liệu character), về mặt thống kê biến này không có ý nghĩa nên sẽ được loại bỏ khỏi dữ liệu khi khảo sát.

- Vì Diff = Mem-Score-Before - Mem-Score-After (đa cộng tuyén), vì vậy ta chỉ cần khảo sát biến phụ thuộc Diff, các biến còn lại sẽ loại bỏ ra khỏi dữ liệu trước khi khảo sát.
- Trên thực tế, việc khảo sát trên từng mức nhóm tuổi trải rộng từ 24 đến 83 rất nhiều, ta sẽ tiến hành chia thành 2 nhóm chính là nhóm tuổi < 50 và nhóm còn lại.
- Biến phụ thuộc là 'Diff' và các biến độc lập bao gồm HappySadgroup, Dosage, Drug và age.

2.1.5. Xử lý dữ liệu

Ở bước này, chúng ta sẽ tiến hành các bước sau:

1. Loai bỏ các biến dư thừa và đưa dữ liệu về dạng thích hợp
2. Xây dựng mô hình AOV để xem xét sự phụ thuộc của biến phụ thuộc vào các biến độc lập. Từ đó chọn ra các biến phù hợp để phân tích.
3. Visualize các biến cần phân tích theo nhóm và rút ra nhận xét.

Sau đây là chi tiết các bước:

- **Bước 1: Loại bỏ các biến cần thiết và asFactor các biến dạng category về dạng factor**

```

1 # Remove unnecessary variables
2 processed_islander = islander_raw[, !(names(islander_raw) %
 3   %in% c("first_name", "last_name", "Mem_Score_Before", "Mem_Score_After"))]
4
5 processed_islander$age = processed_islander$age >= 50
6 processed_islander$age = factor(processed_islander$age)
7 processed_islander$Dosage = factor(processed_islander$Dosage)
8 processed_islander$Drug = factor(processed_islander$Drug)
9 processed_islander$Happy_Sad_group = factor(processed_islander$Happy_Sad_group)
10 levels(processed_islander$age)
11 levels(processed_islander$Drug)
12 levels(processed_islander$Dosage)
13 levels(processed_islander$Happy_Sad_group)

```

Kết quả thu được như sau:

```

1 'FALSE' 'TRUE'
2 'A' 'S' 'T'
3 '1' '2' '3'
4 'H' 'S'
```

Kết quả cho ta thấy rằng nhóm tuổi được chia thành 2 nhóm là trước 50 tuổi và từ 50 tuổi trở về sau, có 3 nhóm thuộc là A, S, T; có 3 liều lượng thuốc được sử dụng là 1, 2, 3 tương ứng với thấp, trung bình và cao; số người khảo sát nằm trong 2 nhóm là Happy và Sad (H, S).

- **Bước 2: Xây dựng mô hình AOV để xem xét sự phụ thuộc của biến phụ thuộc vào các biến độc lập như thế nào**

```

1 diff_aov = aov(Diff ~ ., data = processed_islander)
2 summary(diff_aov)
```

Kết quả

	Df	Sum Sq	Mean Sq	F value	Pr(>F)					
age	1	2	2.1	0.024	0.87821					
Happy_Sad_group	1	11	10.6	0.117	0.73233					
Dosage	2	1222	610.8	6.787	0.00142 **					
Drug	2	4361	2180.6	24.229	4.19e-10 ***					
Residuals	191	17190	90.0							

Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'.'	0.1	' '

Nhìn vào bảng kết quả, ta thấy rằng thực tế các biến ‘HappySadgroup’ và ‘age’ không tham gia vào quá trình giải thích ý nghĩa của biến phụ thuộc ‘Diff’ (với mức ý nghĩa 5%). Do đó, ta chỉ chọn 2 biến ‘Drug’ và ‘Dosage’ để tiến hành khảo sát. Vậy:

- **Mục tiêu:** Khảo sát về tác dụng phụ của các loại thuốc chống trầm cảm đối với trí nhớ của người tham gia thử nghiệm, được đánh giá thông qua thời gian hoàn thành một bài kiểm tra trí nhớ
- **Biến phản hồi:** ‘Diff’ cho biết chênh lệch giữa thời gian (giây) hoàn thành bài kiểm tra trước và sau khi sử dụng thuốc.
- **Biến nhân tố: Drug:** Gồm 3 nhóm ‘A’ (Alprazolam), ‘S’ (Placebo) và ‘T’ (Triazolam) và **Dosage:** Gồm 3 nhóm ‘1’ (thấp), ‘2’ (trung bình) và ‘3’ (Cao)

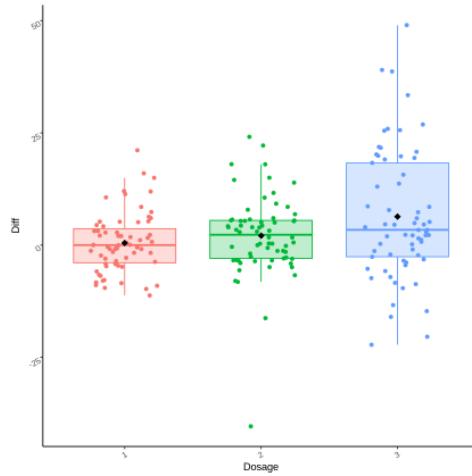
- **Bước 3: Visualize dữ liệu của các biến theo từng nhóm**

```

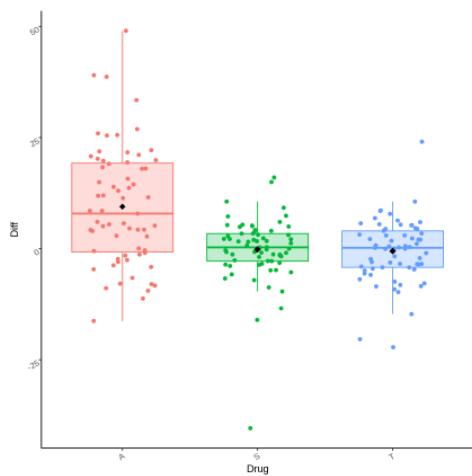
1 processed_islander = processed_islander[, !(names(processed
2   _islander) %in% c("age", "Happy_Sad_group"))]
3 # Dosage variable
4 ggplot(processed_islander ,aes(x=Dosage, y=Diff, colour=
5   Dosage, fill=Dosage))+ 
6   geom_jitter(width=0.25)+ 
7   geom_boxplot(alpha=0.25, outlier.alpha=0) + 
8   stat_summary(fun.y=mean, colour="black", geom="point",
9     shape=18, size=3, show.legend = FALSE) + 
10  theme_classic() +
11  theme(legend.position="none")+
12  theme(axis.text = element_text(angle=30, hjust=1, vjust
13    =1))
14
15 # Drug variable
16 ggplot(processed_islander ,aes(x=Drug, y=Diff, colour=Drug,
17   fill=Drug))+ 
18   geom_jitter(width=0.25)+ 
19   geom_boxplot(alpha=0.25, outlier.alpha=0) + 
20   stat_summary(fun.y=mean, colour="black", geom="point",
21     shape=18, size=3, show.legend = FALSE)+ 
22  theme_classic()+
23  theme(legend.position="none")+
24  theme(axis.text = element_text(angle=30, hjust=1, vjust
25    =1))

```

Kết quả ta thu được như sau:



Hình 2.4: Phân phối dữ liệu của từng nhóm Dosage



Hình 2.5: Phân phối dữ liệu của từng nhóm Drug

Nhìn vào đây, ta có thể rút ra một vài nhận xét như sau:

– Đối với Dosage

- * Liều lượng thấp: Trung vị Diff gần bằng 0, với phạm vi nhỏ. Hầu hết các điểm dữ liệu tập trung xung quanh trung vị, với một vài ngoại lệ.
- * Liều lượng trung bình: Trung vị Diff vẫn gần bằng 0, nhưng phạm vi dữ liệu lớn hơn một chút so với liều lượng thấp. Có sự phân bố rộng hơn của các điểm, cho thấy sự biến đổi nhiều hơn trong các phản ứng.
- * Liều lượng cao: Trung vị Diff cao hơn so với các nhóm khác, gợi ý rằng liều lượng thuốc này có tác dụng lớn hơn. Có sự biến đổi lớn, với một số điểm nằm khá cao hoặc thấp so với trung vị. IQR lớn hơn và sự hiện diện của các ngoại lệ cho thấy một phạm vi rộng của các phản ứng đối với liều lượng cao.

Ta rút ra kết luận như sau:

- * Biểu đồ gợi ý rằng liều lượng cao hơn có thể có tác động đáng kể hơn đến thời gian hoàn thành bài kiểm tra trí nhớ, được chỉ ra bởi trung vị cao hơn và sự biến đổi lớn hơn trong Diff.
- * Liều lượng thấp và trung bình cho thấy sự thay đổi nhỏ hơn và ít biến đổi hơn, với nhiều người tham gia có ít hoặc không thay đổi trong thời gian hoàn thành.
- * Sự phân bố rộng hơn và trung vị cao hơn trong nhóm liều lượng cao có thể cho thấy rằng mặc dù một số người tham gia có lợi ích đáng kể, những người khác có thể gặp tác dụng phụ, dẫn đến phản ứng không tốt.

– Đối với Drug

- * Alprazolam (A): Trung vị Diff khá xa với điểm 0 hơn các loại thuốc khác. Có nhiều điểm ngoại lệ, đặc biệt ở phía trên, cho thấy một số người tham gia có sự cải thiện đáng kể về thời gian hoàn thành bài kiểm tra. Trung bình Diff cũng cao hơn, gợi ý rằng Alprazolam có thể có tác dụng tích cực đối với một số người tham gia.
- * Triazolam (T): Trung vị Diff gần bằng 0, với phạm vi và sự phân tán nhỏ hơn so với Alprazolam. Trung bình Diff gần với trung vị, cho thấy tác dụng của Triazolam ít biến đổi hơn.
- * Placebo (S): Trung vị Diff gần bằng 0, với phạm vi và sự phân tán tương tự như Triazolam. Có một số điểm ngoại lệ, nhưng không nhiều. Trung bình Diff gần với trung vị, cho thấy tác dụng của giả dược (Placebo) ít biến đổi và không có hiệu quả đáng kể.

Ta rút ra kết luận như sau:

- * Alprazolam (A): Có tác dụng đáng kể đối với một số người tham gia, nhưng cũng có nhiều biến đổi, cho thấy có thể có tác dụng phụ hoặc tác động không đồng nhất.
- * Triazolam (T) và Placebo (S): Không có sự thay đổi đáng kể trong thời gian hoàn thành bài kiểm tra, gợi ý rằng các loại thuốc này có ít hoặc không có tác dụng cải thiện trí nhớ.

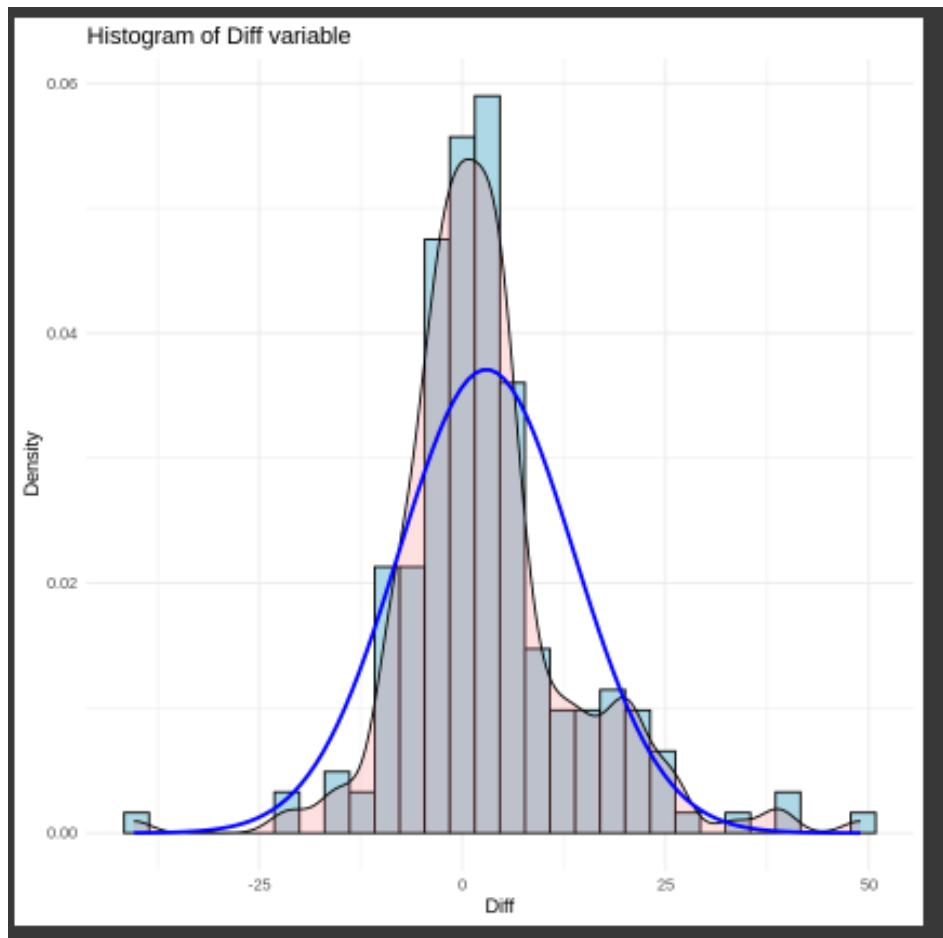
Ta xem xét biến phụ thuộc Diff bằng cách xét lại biểu đồ phân bố chuẩn như sau:

```
1 ggplot(processed_islander, aes(x = Diff)) +  
2   geom_histogram(aes(y = ..density..), bins = 30, color = "black", fill = "lightblue") +  
3   geom_density(alpha = 0.2, fill = "#FF6666") +  
4   stat_function(fun = dnorm, args = list(mean = mean(  
5     processed_islander$Diff, na.rm = TRUE), sd = sd(  
6     processed_islander$Diff, na.rm = TRUE)),  
7                 color = "blue", size = 1) +  
8   theme_minimal()
```

```

7  labs(title = "Histogram of Diff variable", x = "Diff", y
      = "Density")

```



Hình 2.6: Đồ thị phân phối chuẩn của biến ‘Diff’

Ta cũng rút ra một vài nhận xét như sau:

- Độ lệch
 - * Trung bình (Mean) của biến Diff là 2.955, cho thấy thời gian trung bình hoàn thành bài kiểm tra sau khi dùng thuốc tăng thêm khoảng 2.955 giây.
 - * Median (Trung vị) là 1.700, thấp hơn giá trị trung bình, cho thấy sự phân phối không hoàn toàn đối xứng.
- Phân Bố Dữ Liệu: Dữ liệu phân bố khá gần với phân phối chuẩn, nhưng có một vài khác biệt:
 - * Có một sự tập trung dữ liệu khá cao xung quanh giá trị 0 đến 5, tạo nên một đỉnh phân phối cao hơn so với đường chuẩn.

- * Có sự lan tỏa dữ liệu về cả hai phía của đỉnh, nhưng ít hơn ở phía cực trái (khoảng -40) và cực phải (khoảng 50).
- Khoảng Tứ Phân Vị:
 - * 1st Qu. (Quartile đầu tiên): -3.175.
 - * 3rd Qu. (Quartile thứ ba): 5.925.
 - * Điều này cho thấy phần lớn dữ liệu nằm trong khoảng từ -3.175 đến 5.925, một khoảng phân tán rộng nhưng không đối xứng hoàn toàn quanh trung vị.
- Đỉnh và đuôi: Biểu đồ cho thấy một đỉnh lớn, khá hẹp, và hai đuôi phân phối khá dài, đặc biệt ở phía phải (hơn 25). Điều này chỉ ra rằng có một số giá trị cực đại cao (tăng lớn trong thời gian hoàn thành bài kiểm tra sau khi dùng thuốc).

Kết Luận:

- Sự Phân Tán và Độ Lệch: Biểu đồ cho thấy sự tăng thời gian hoàn thành bài kiểm tra (Diff) sau khi dùng thuốc là phổ biến, với giá trị trung bình dương. Tuy nhiên, phân phối không hoàn toàn đối xứng, với một số giá trị cực đại ở cả hai phía.
- Khả Năng Phân Phối Chuẩn: Phân phối của biến Diff khá gần với phân phối chuẩn, nhưng có một số khác biệt như đỉnh phân phối cao hơn và đuôi phân phối dài hơn. Điều này có thể là do tác động của một số cá nhân phản ứng mạnh mẽ với thuốc hơn so với phần còn lại.

2.1.6. Kiểm định các giả thiết thống kê (ANOVA assumptions)

Nhắc lại các điều kiện để phân tích ANOVA như sau:

1. Các mẫu độc lập
2. Biến phụ thuộc là biến liên tục
3. Các nhóm có phân phối chuẩn hoặc gần chuẩn, đồng nghĩa với việc kiểm định phương sai các nhóm cho kết quả là đồng nhất.

Rõ ràng, theo như phân tích phía trên, bộ dữ liệu chúng ta đã thỏa mãn điều kiện (1) và (2). Để chắc chắn ta sẽ đi kiểm định yêu cầu số (3) bằng cách tiến hành thực hiện các kiểm định sau:

- Shapiro-Wilk test
- leveneTest
- durbinWatsonTest

Đầu tiên ta sẽ xây dựng mô hình tương tác bằng dòng lệnh sau

```
1 # Xây dựng mô hình tương tác
2 int_model = aov(Diff~Dosage * Drug, data = processed_islander)
3 summary(int_model)
```

Kết quả thu được:

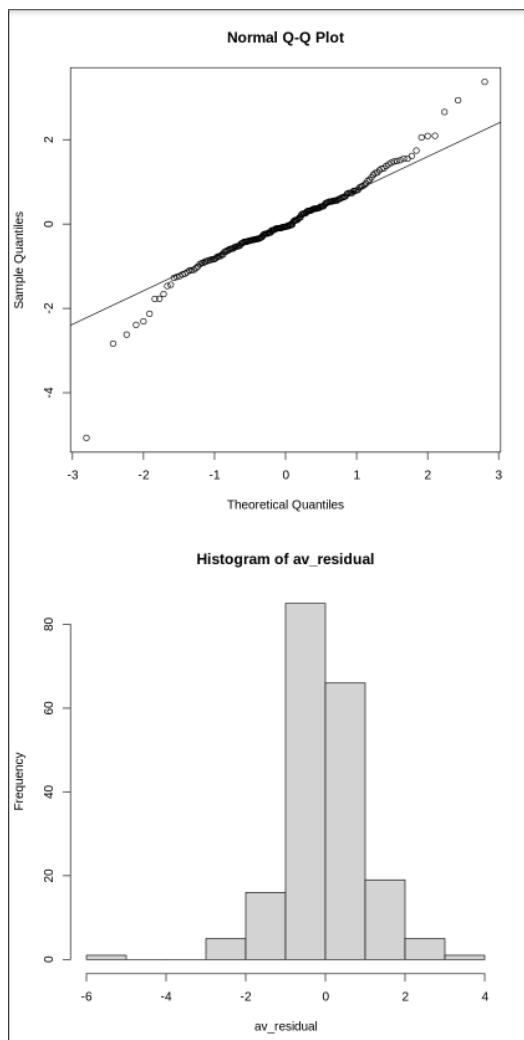
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
1 Dosage	2	1222	610.9	9.536	0.000113 ***	
2 Drug	2	4314	2156.9	33.666	3.12e-13 ***	
3 Dosage:Drug	4	5141	1285.3	20.062	8.74e-14 ***	
4 Residuals	189	12109	64.1			
5 ---						
6 Signif. codes:	0	'***'	0.001	'**'	0.01	'*'
					0.05	'.'
					0.1	''
					1	

Với mức ý nghĩa 5%, ta thấy rằng giữa ‘Dosage’ và ‘Drug’ có mối quan hệ tương tác với nhau dẫn đến tác động hiệu quả của việc dùng thuốc đối với trí nhớ của người sử dụng. Tiếp tục đi kiểm định các thông số sau:

```

1 # Shapiro-Wilk test
2 int_model = aov(Diff~Dosage * Drug, data = processed_islander)
3 av_residual = rstandard(int_model)
4 shapiro.test(av_residual)
5 # QQ plot
6 qqnorm(av_residual)
7 qqline(av_residual)
8 hist(av_residual)

```



Hình 2.7: Biểu đồ phần dư

Kết quả như sau:

```
1      Shapiro-Wilk normality test
2
3 data: av_residual
4 W = 0.95575, p-value = 7.794e-06
```

Xét giả định

- H0: Tuân theo phân phối chuẩn
- H1: Không tuân theo phân phối chuẩn

Nhận xét: Với độ tin cậy 5% thì với giá trị p-value = 7.794e-06 chúng ta đủ cơ sở bác bỏ H0, vậy phần dư có phân phối không chuẩn. Nhìn vào biểu đồ, ta thấy rằng ở phần đuôi kéo dài, có một vài điểm bị kéo lệch ra khỏi đường thẳng, điều này chứng tỏ rằng khả năng các điểm nhiễu chính là các điểm ngoại lệ (outliers), tuy nhiên, về mặt tổng quan, dữ liệu vẫn có dạng gần chuẩn, do đó, ta vẫn có thể tiến hành kiểm tra ANOVA. Ở bước cải tiến, ta sẽ tiến hành xử lý các điểm ngoại lệ này để so sánh với kiểm định ban đầu.

Bước tiếp theo ta Kiểm định các nhóm có phương sai đồng nhất hay không

```
1 leveneTest(int_model)
```

Kết quả

```
1 Df          F value  Pr(>F)
2 <int>        <dbl>    <dbl>
3 group       8        0.9961061   0.4404991
4 189        NA       NA
```

Giả định:

- H0: Các nhóm có phương sai đồng nhất
- H1: Các nhóm không có phương sai đồng nhất

Nhận xét: Với giá trị p-value = 0.4404991 > 0.05, ta không đủ điều kiện bác bỏ H0, vậy các nhóm có phương sai đồng nhất. Như vậy, bộ dữ liệu này đã thoả mãn điều kiện yêu cầu để phân tích ANOVA. Tuy nhiên, để hiểu sâu hơn về bộ dữ liệu, ta tiếp tục đi kiểm định tính độc lập của phần dư bằng dòng lệnh sau

```

1 # Kiểm định tính độc lập của phần dư
2 durbinWatsonTest(int_model)
3 plot(int_model, 1)

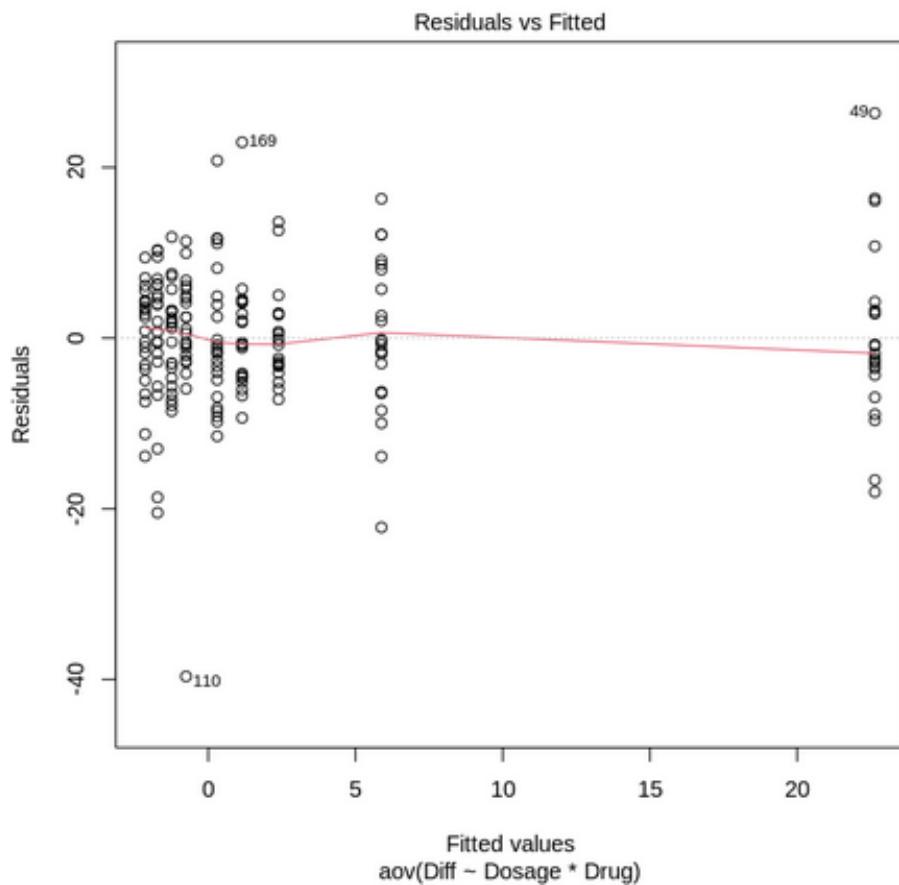
```

Kết quả

```

1 lag Autocorrelation D-W Statistic p-value
2   1      -0.05315976      2.105716    0.816
3 Alternative hypothesis: rho != 0

```



Hình 2.8: Kiểm định độc lập phần dư

Với giả định:

- H₀: Không có sự tương quan (độc lập).
- H₁: H₁: Có sự tương quan (không độc lập).

Thì với giá trị p-value = 0.816 (> 0.05) nên không có sự tương quan. Vậy phần dư độc lập

2.1.7. Phân tích phương sai k nhân tố

Tiếp theo chung ta sẽ tiến hành đi phân tích phương sai k nhân tố. Việc này gồm các bước sau:

1. Kiểm tra sự tương tác
2. Phân tích ảnh hưởng đơn
 - Phân tích ảnh hưởng đơn của liều lượng ở mỗi loại thuốc
 - Phân tích ảnh hưởng đơn của thuốc ở mỗi liều lượng
3. Phân tích ảnh hưởng chính
 - Phân tích ảnh hưởng chính của Dosage với hiệu quả của bài kiểm tra trí nhớ
 - Phân tích ảnh hưởng chính của Drug với hiệu quả của bài kiểm tra trí nhớ

Sau đây là các bước chi tiết:

- **Bước 1: Xây dựng mô hình tương tác (interaction model) và kiểm tra tương tác của các biến**

```
1 int_model = aov(Diff~Dosage * Drug, data = processed_islander)
2 summary(int_model)
3 plot(interactionMeans(int_model))
```

Kết quả

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
1 Dosage	2	1222	610.9	9.536	0.000113	***
2 Drug	2	4314	2156.9	33.666	3.12e-13	***
3 Dosage:Drug	4	5141	1285.3	20.062	8.74e-14	***
4 Residuals	189	12109	64.1			
5 ---						
6 Signif. codes:	0	'***'	0.001	'**'	0.01	'*'
					0.05	'.'
					0.1	' '
						1

Ta rút ra một số nhận xét dựa trên kết quả như sau:

- Với mức ý nghĩa 5%, ta thấy giữa ‘Dosage’ và ‘Drug’ có sự tương tác với nhau ($p\text{-value}=8.74\text{e}-14$). Sự kết hợp giữa hàm lượng thuốc và loại thuốc có ảnh hưởng rất lớn đến thời gian hoàn thành bài kiểm tra, cho thấy rằng không chỉ từng yếu tố riêng lẻ mà sự kết hợp giữa chúng cũng rất quan trọng.
- Đối với biểu đồ bên trái: Biểu đồ bên trái biểu diễn sự tương tác giữa Dosage và Drug dựa trên giá trị trung bình điều chỉnh của Diff.

+ Đồ thị trên cùng bên trái (Dosage theo Drug):

- * Cho thấy sự thay đổi của Diff theo liều lượng (Dosage) cho từng loại thuốc (Drug).
- * Đối với tất cả các loại thuốc, Diff tăng dần khi tăng liều lượng từ 1 đến 3.

+ Đồ thị dưới cùng bên trái (Drug theo Dosage):

- * Cho thấy sự thay đổi của Diff theo loại thuốc (Drug) cho từng liều lượng (Dosage).
- * Khi liều lượng là 1: sự khác biệt giữa các loại thuốc là tương đối nhỏ. Trong đó loại S cho hiệu quả cao nhất, A và T cho kết quả tệ hơn trước khi sử dụng (mean < 0).
- * Khi liều là 2: Có sự thay đổi rõ rệt ở các loại thuốc: Loại S cho kết quả tệ hơn so với trước khi dùng thuốc, loại T có hiệu quả không đáng kể, loại A cho thấy hiệu quả vượt bật.
- * Khi liều lượng là 3: Sự khác biệt giữa các loại thuốc trở nên rõ rệt hơn, với Thuốc A cho kết quả tốt nhất so với 3 loại và ở cả 3 liều lượng, trong khi 2 loại còn lại cho kết quả tệ hơn trước khi dùng.

- Đối với biểu đồ bên phải: Biểu đồ bên phải biểu diễn sự tương tác giữa Drug và Dosage dựa trên giá trị trung bình điều chỉnh của Diff.

+ Đồ thị trên cùng bên phải (Drug theo Dosage):

- * Tương tự như đồ thị dưới cùng bên trái của biểu đồ bên trái, nhưng theo chiều ngược lại. Cho thấy sự thay đổi của Diff theo loại thuốc cho từng liều lượng.
- * Đối với thuốc A: Cho kết quả tốt liều lượng cao và trung bình, liều lượng thấp không có sự thay đổi
- * Đối với thuốc T: Không có sự khác biệt giữa các liều lượng và hiệu quả sau và trước khi sử dụng (Thậm chí giảm (tệ))
- * Đối với thuốc S: Giống T

+ Đồ thị dưới cùng bên phải (Dosage theo Drug):

- * Tương tự như đồ thị trên cùng bên trái của biểu đồ bên trái, nhưng theo chiều ngược lại.
- * Cho thấy sự thay đổi của Diff theo liều lượng cho từng loại thuốc.
- * Cơ bản các thuốc cho kết quả tốt nhất theo thứ tự là A > S > T.

Ta có kết luận sau đây:

- Cả liều lượng và loại thuốc đều có ảnh hưởng đáng kể đến Diff.
- Có sự tương tác đáng kể giữa liều lượng và loại thuốc, nghĩa là hiệu quả của liều lượng khác nhau phụ thuộc vào loại thuốc được sử dụng.
- Biểu đồ cho thấy rõ ràng rằng Thuốc A (Alprazolam) có ảnh hưởng lớn nhất khi liều lượng tăng, trong khi S có ảnh hưởng ít nhất.

- **Bước 2: Phân tích ảnh hưởng đơn**

Để phân tích ảnh hưởng đơn ta sẽ sử dụng hàm **testInteractions** để tiến hành phân tích. Sau đây là các bước chi tiết

- Phân tích ảnh hưởng đơn của liều lượng ở mỗi loại thuốc

```
1 testInteractions(int_model, fixed = "Drug", across = "
    Dosage")
```

Kết quả:

Warning message in rbind(deparse.level, ...): "number of columns of result, 8, is not a multiple of vector length 6 of arg 2" A anova: 4 x 8								
	Dosage1	Dosage2	SE1	SE2	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
A	-22.3365613	-16.759091	2.386970	2.413346	2	6031.8648	47.0749202	7.712930e-17
S	4.5318182	1.390909	2.413346	2.413346	2	237.1403	1.8507313	3.199167e-01
T	0.4829004	2.873810	2.441908	2.441908	2	102.9239	0.8032563	4.493911e-01
Residuals	NA	NA	189.000000	12108.596751	NA	NA	NA	NA

Hình 2.9: Kết quả ảnh hưởng đơn giữa Drug và Dosage

Với các giả định như sau:

- * H0: Liều lượng không ảnh hưởng đến hiệu quả thuốc
- * H1: Liều lượng có ảnh hưởng đến hiệu quả của thuốc

Ta rút ra kết luận như sau: Với kết quả phân tích ta có một số nhận xét như sau, với độ tin cậy 5% thì:

- * Liều lượng có ảnh hưởng đến kết quả của loại thuốc A
- * Liều lượng không ảnh hưởng đến kết quả của lô thuốc S và T

- Phân tích ảnh hưởng đơn của thuốc ở mỗi liều lượng

```
1 testInteractions(int_model, fixed = "Dosage",
    across = "Drug")
```

Warning message in rbind(deparse.level, ...):								
"number of columns of result, 8, is not a multiple of vector length 6 of arg 2"								
Anova: 4 x 8								
	Drug1	Drug2	SE1	SE2	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1.545257	3.627273	2.386970	2.413346	2	145.8161	1.138003	3.226440e-01
2	4.731818	-1.904545	2.413346	2.413346	2	513.7639	4.009605	3.941034e-02
3	24.364719	-0.421645	2.441908	2.441908	2	8795.3182	68.641940	1.169435e-22
Residuals	NA	NA	189.000000	12108.596751	NA	NA	NA	NA
Tukey multiple comparisons of means 95% family-wise confidence level								

Hình 2.10: Ánh hưởng đơn giữa Drug và Dosage.

Tương tự như ở phía trên, ta có các giả định như sau:

- * H0: Các loại thuốc sẽ không tác động ở mỗi liều lượng
- * H1: Các loại thuốc sẽ có tác động ở mỗi liều lượng

Ta rút ra kết luận như sau: Với kết quả phân tích ta có một số nhận xét như sau, với độ tin cậy 5% thì:

- * Hầu hết các loại thuốc sẽ có tác động ở liều lượng cao
 - * Liều lượng trung bình cũng sẽ có tác động nhưng không đáng kể (có ý nghĩa ở mức 0.4)
 - * Liều lượng thấp cho kết quả không đáng kể.
- Phân tích ảnh hưởng đơn giữa các nhóm thuốc ứng với mỗi liều lượng

Việc phân tích sự tương tác của các nhóm trong cùng một liều lượng cũng có ý nghĩa rất quan trọng trong thống kê, từ đó sẽ hiểu rõ hơn về từng tác dụng của từng loại và từng nhóm

```

1 options(contrasts = c(unordered="contr.sum", ordered=
2   "contr.poly"))
3 A_vs_S = list(Drug = c(1, -1, 0))
4 A_vs_T = list(Drug = c(1, 0, -1))
5 S_vs_T = list(Drug = c(0, 1, -1))

```

Đầu tiên, ta sẽ đi phân tích ảnh hưởng của nhóm A và S

```

1 testInteractions(int_model, custom = c(A_vs_S), fixed =
2   "Dosage", adjustment = "bonferroni")

```

Kết quả:

"number of columns of result, 6, is not a multiple of vector length 5 of arg 2"						
A anova: 4 × 6						
	Value	SE	Df	Sum of Sq	F	Pr (>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1 : Drug1	-2.082016	2.386970	1.0	48.7423	0.7608062	1.000000e+00
2 : Drug1	6.636364	2.413346	1.0	484.4545	7.5617275	1.962325e-02
3 : Drug1	24.786364	2.413346	1.0	6758.0020	105.4839312	1.813548e-19
Residuals	NA	189.000000	12108.6	NA	NA	NA

Hình 2.11: Tương tác giữa nhóm A và S ở mỗi liều lượng

Ta có giả định như sau:

- * Không có sự khác nhau giữa nhóm thuốc A và S ở các liều lượng
- * Có sự khác nhau giữa nhóm thuốc A và S ở các liều lượng

Ta rút ra kết luận như sau: Với kết quả phân tích ta có một số nhận xét như sau, với độ tin cậy 5% thì:

- * Có sự khác biệt về hiệu quả khi sử dụng thuốc thuốc A và S ở các liều lượng cao và trung bình
- * Ở liều lượng thấp: Không có sự khác biệt

Tiếp theo là nhóm A và T

```
1 testInteractions(int_model, custom = c(A_vs_T), fixed =
  "Dosage", adjustment = "bonferroni")
```

Kết quả:

"number of columns of result, 6, is not a multiple of vector length 5 of arg 2"						
A anova: 4 × 6						
	Value	SE	Df	Sum of Sq	F	Pr (>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1 : Drug1	1.545257	2.386970	1.0	26.8497	0.4190901	1.000000e+00
2 : Drug1	4.731818	2.413346	1.0	246.2911	3.8442956	1.541547e-01
3 : Drug1	24.364719	2.441908	1.0	6378.1734	99.5552823	1.262176e-18
Residuals	NA	189.000000	12108.6	NA	NA	NA

Hình 2.12: Tương tác giữa nhóm A và T ở mỗi liều lượng

Với các giả định tương tự với nhóm A và S, ta có kết luận như sau:

- * Ở liều cao: Có sự khác biệt về hiệu quả khi sử dụng thuốc ở các loại thuốc A và T
- * Ở liều thấp và trung bình: Không có sự ảnh hưởng rõ rệt

Cuối cùng là giữa nhóm S và T

```
1 testInteractions(int_model, custom = c(S_vs_T), fixed =
    "Dosage", adjustment = "bonferroni")
```

Kết quả:

A anova: 4 x 6						
	Value	SE	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1 : Drug1	3.627273	2.413346	1.0	144.72818	2.25902530	0.4035201
2 : Drug1	-1.904545	2.413346	1.0	39.90023	0.62279248	1.0000000
3 : Drug1	-0.421645	2.441908	1.0	1.91015	0.02981504	1.0000000
Residuals	NA	189.000000	12108.6	NA	NA	NA

Hình 2.13: Tương tác giữa nhóm S và T ở mỗi liều lượng

Với các giả định tương tự với nhóm A và S, ta có kết luận như sau: Với độ tin cậy 5% thì không có sự khác biệt nào ở cả 3 liều lượng.

Từ việc phân tích trên, ta có kết luận như sau: Khi dùng thuốc liều cao, giữa các nhóm sẽ cho ra các phản ứng ảnh hưởng đến trí nhớ ở nhóm A-T và A-S. Tuy nhiên, nhóm S-T lại không cho thấy sự tương tác nào có ý nghĩa thống kê.

- Phân tích ảnh hưởng đơn giữa các nhóm liều lượng ứng với mỗi loại thuốc Tương tự cho trường hợp **Phân tích ảnh hưởng đơn giữa các nhóm thuốc ứng với mỗi liều lượng**, ta cũng phân tích ảnh hưởng đơn giữa các nhóm liều lượng ứng với mỗi loại thuốc như thế nào

```
1 options(contrasts = c(unordered="contr.sum", ordered="contr
    .poly"))
2 low_vs_medium = list(Dosage = c(1, -1, 0))
3 low_vs_high = list(Dosage = c(1, 0, -1))
4 medium_vs_high = list(Dosage = c(0, 1, -1))
```

Đầu tiên là nhóm thấp và trung bình

```
1 # Nhóm thấp và trung bình
2 testInteractions(int_model, custom = c(low_vs_medium),
    fixed = "Drug", adjustment = "bonferroni")
```

Kết quả:

"number of columns of result, 6, is not a multiple of vector length 5 of arg 2"						
A anova: 4 x 6						
	Value	SE	Df	Sum of Sq	F	Pr (>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
A : Dosage1	-5.577470	2.386970	1.0	349.79415	5.4598477	0.06151837
S : Dosage1	3.140909	2.413346	1.0	108.51841	1.6938362	0.58404073
T : Dosage1	-2.390909	2.413346	1.0	62.88091	0.9814921	0.96929207
Residuals	NA	189.000000	12108.6	NA	NA	NA

Hình 2.14: Tương tác giữa nhóm thấp và nhóm trung bình

Với giả định sau:

- H0: Không có sự tương tác nhau giữa liều thấp và liều trung bình ở loại thuốc X (A, T, S)
- H1: Có sự tương tác nhau giữa liều thấp và liều trung bình ở loại thuốc X (A, T, S)

Với độ tin cậy 5% thì

- Ở loại thuốc A: Có sự tương tác về hiệu quả khi sử dụng thuốc ở các liều lượng thấp và liều lượng trung bình
- Ở loại thuốc S và T: Không có sự tương tác

Tiếp theo là nhóm thấp và cao

```
1 # Nhóm thấp và cao
2 testInteractions(int_model, custom = c(low_vs_high), fixed
= "Drug", adjustment = "none")
```

"number of columns of result, 6, is not a multiple of vector length 5 of arg 2"						
A anova: 4 x 6						
	Value	SE	Df	Sum of Sq	F	Pr (>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
A : Dosage1	-16.759091	2.413346	1.0	3089.53841	48.2238175	5.939718e-11
S : Dosage1	1.390909	2.413346	1.0	21.28091	0.3321683	5.650708e-01
T : Dosage1	2.873810	2.441908	1.0	88.73388	1.3850245	2.407272e-01
Residuals	NA	189.000000	12108.6	NA	NA	NA

Hình 2.15: Tương tác giữa nhóm thấp và cao

Tương tự như trên ta rút ra nhận xét như sau: Với mức ý nghĩa 5%

- Ở loại thuốc A: Có sự tương tác khi sử dụng thuốc ở các liều lượng thấp và liều lượng cao
- Ở loại thuốc S và T: Không có sự tương tác mang ý nghĩa thống kê

Cuối cùng là nhóm trung bình và cao

```
1 # Nhóm trung bình và cao
2 testInteractions(int_model, custom = c(medium_vs_high),
fixed = "Drug", adjustment = "none")
```

Warning message in rbind(deparse.level, ...): "number of columns of result, 6, is not a multiple of vector length 5 of arg 2" Anova: 4 x 6						
	Value	SE	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
A : Dosage1	-16.759091	2.413346	1.0	3089.53841	48.2238175	5.939718e-11
S : Dosage1	1.390909	2.413346	1.0	21.28091	0.3321683	5.650708e-01
T : Dosage1	2.873810	2.441908	1.0	88.73388	1.3850245	2.407272e-01
Residuals	NA	189.000000	12108.6	NA	NA	NA

Hình 2.16: Tương tác giữa nhóm trung bình và cao

Tương tự như trên ta rút ra nhận xét như sau: Với mức ý nghĩa 5%

- Có sự tương tác khi sử dụng thuốc ở các liều lượng trung bình và liều lượng cao
- Không có sự tương tác mang ý nghĩa thống kê

Qua việc phân tích trên ta có một số kết luận như sau:

- Hầu hết các liều lượng thấp và trung bình sẽ cho thấy mức độ ít tương tác mang ý nghĩa thống kê ở các loại thuốc
- Hầu hết các loại thuốc T và S cho kết quả là không có sự tương tác giữa liều lượng với loại A

• **Bước 3: Phân tích ảnh hưởng chính**

Ở bước này ta sẽ thực hiện 2 phân tích:

- Phân tích ảnh hưởng chính của Dosage với hiệu quả của bài kiểm tra trí nhớ
- Phân tích ảnh hưởng chính của Drug với hiệu quả của bài kiểm tra trí nhớ

Với mỗi bước, ta sẽ thực hiện các công việc sau:

- Xây dựng mô hình
- Kiểm định các giả thiết của mô hình
- Kiểm định trung bình của các nhóm
- Nhận xét

Sau đây là các bước phân tích cụ thể:

- **Bước 3.1: Phân tích ảnh hưởng chính của Dosage với hiệu quả của bài kiểm tra trí nhớ**

```
1 dosage_model = aov(Diff~Dosage, data = processed_islander)
2 summary(dosage_model)
```

Kết quả:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
1 Dosage	2	1222	610.9	5.524	0.00464 **
2 Residuals	195	21563	110.6		
3 ---					
4 Signif. codes:	0	'***'	0.001	'**'	0.01
	0.05	'.'			
	0.1	' '	1		

Nhận xét: Với mức ý nghĩa 0.05, ta thấy rằng Dosage có ý nghĩa trong việc giải thích mô hình

Tiếp theo tiến hành kiểm định các giả thuyết

```
1 # Shapiro-Wilk test
2 av_residual = rstandard(dosage_model)
3 shapiro.test(av_residual)
4
5 # Trực quan bằng QQ plot
6 qqnorm(av_residual)
7 qqline(av_residual)
8 hist(av_residual)
```

Kết quả:

```
1 Shapiro-Wilk normality test
2
3 data: av_residual
4 W = 0.94813, p-value = 1.397e-06
```

Với giả định:

- * H0: Phần dư Tuân theo phân phối chuẩn.
- * H1: Phần dư Không tuân theo phân phối chuẩn.

Như vậy với độ tin cậy 5% thì với giá trị p-value = 1.397e-06 chúng ta đủ cơ sở bác bỏ H₀, vậy sai số có phân phối không chuẩn. Nhìn vào biểu đồ, ta thấy rằng ở phần đuôi kéo dài, có một vài điểm bị kéo lệch ra khỏi đường thẳng \rightarrow Khả năng các điểm nhiễu chính là các điểm ngoại lệ (outliners), tuy nhiên, về mặt tổng quan, dữ liệu vẫn có dạng gần chuẩn.

```
1 # Kiểm định các nhóm có phương sai đồng nhất hay không
2 leveneTest(dosage_model)
```

Kết quả:

```
1          A anova: 2 x 3
2          Df      F value Pr(>F)
3          <int>    <dbl>     <dbl>
4 group     2        11.76277   1.502826e-05
5                 195             NA                  NA
```

Với giả định:

- * Các nhóm có phương sai đồng nhất.
- * Các nhóm không có phương sai đồng nhất.

Nhận xét: Với giá trị p-value = 1.502826e-05 < 0.05 , ta đủ điều kiện bác bỏ H₀, vậy các nhóm có phương sai không đồng nhất.

```
1 # Kiểm định tính độc lập của phần dư
2 durbinWatsonTest(dosage_model)
3 plot(dosage_model, 1)
```

Kết quả:

```
1 lag Autocorrelation D-W Statistic p-value
2   1       0.3993242      1.198467      0
3 Alternative hypothesis: rho != 0
```

Với giả định:

- H₀: Không có sự tương quan (độc lập)
- H₁: Có sự tương quan (không độc lập)

Nhận xét: Với giá trị p-value = 0 nên có sự tương quan dương.

Mặc dù với điều kiện phương sai giữa các nhóm không đồng nhất nên sẽ không tiến hành phân tích ANOVA được, tuy nhiên về mặc trực quan hóa dữ liệu, ta thấy rằng đồ thị phân bố dạng gần chuẩn, nên ta sẽ tiếp tục đi phân tích các yếu tố ANOVA.

```

1 # Kiểm định trung bình giữa các nhóm liều lượng
2 with(processed_islander, pairwise.t.test(Diff, Dosage, p.
   adj = "bonferroni"))
3 TukeyHSD(aov(Diff~Dosage, data=processed_islander), conf.
   level = 0.95)
4 plot(TukeyHSD(aov(Diff~Dosage, data=processed_islander),
   conf.level = 0.95))

```

Kết quả:

```

1
2      Pairwise comparisons using t tests with pooled SD
3
4 data: Diff and Dosage
5
6    1     2
7 2 1.0000 -
8 3 0.0045 0.0620
9
10 P value adjustment method: bonferroni
11
12 Tukey multiple comparisons of means
13   95% family-wise confidence level
14
15 Fit: aov(formula = Diff ~ Dosage, data = processed_islander
16
17 $Dosage
18
19      diff        lwr         upr      p adj
20 2-1  1.611827 -2.69534819  5.919003 0.6511582
21 3-1  5.899403  1.57556917 10.223237 0.0042357
22 3-2  4.287576 -0.05235809  8.627510 0.0536311

```

Với giả định:

- H0: Các giá trị trung bình giữa các cặp bằng nhau
- H1: Các giá trị trung bình giữa các cặp không bằng nhau

Nhìn vào kết quả ta có: Cặp 3-1 có p-value đều có giá trị nhỏ hơn 0.05 (độ tin cậy 95%) nên ta cơ sở để bác bỏ H0. Vậy rõ ràng giữa các nhóm này có giá trị trung bình là khác nhau. Nghĩa

là các nhóm thuốc liều cao và thấp thì cho thấy mức độ ảnh hưởng đến bệnh nhân khác nhau. Còn các nhóm còn lại thì không, Để rõ hơn, nhìn vào kết quả và hình vẽ ta cũng thấy ngay giữa nhóm 3-1 có mức độ hiệu quả trung bình khác nhau, 3-2 và 1-2 có mức độ hiệu quả trung bình như nhau (đồ thị cắt điểm 0).

```
1 # Phân tích tương tác của từng nhóm liều lượng với nhau
2 A_vs_S = list(Dosage = c(1, -1, 0))
3 A_vs_T = list(Dosage = c(1, 0, -1))
4 S_vs_T = list(Dosage = c(0, 1, -1))
5 testInteractions(dosage_model, custom = A_vs_S, adjustment
                  = 'bonferroni')
6 print("-----")
7 testInteractions(dosage_model, custom = A_vs_T, adjustment
                  = 'bonferroni')
8 print("-----")
9 testInteractions(dosage_model, custom = S_vs_T, adjustment
                  = 'bonferroni')
```

Kết quả

```

Warning message in rbind(deparse.level, ...):
"number of columns of result, 6, is not a multiple of vector length 5 of arg 2"
A anova: 2 × 6
  Value        SE      Df Sum of Sq       F     Pr(>F)
<dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
Dosage1 -1.611827 1.823722 1.00 86.37818 0.7811232 0.3778862
Residuals NA 195.000000 21563.49 NA NA NA
[1] -----
Warning message in rbind(deparse.level, ...):
"number of columns of result, 6, is not a multiple of vector length 5 of arg 2"
A anova: 2 × 6
  Value        SE      Df Sum of Sq       F     Pr(>F)
<dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
Dosage1 -5.899403 1.830776 1.00 1148.234 10.38355 0.001490166
Residuals NA 195.000000 21563.49 NA NA NA
[1] -----
Warning message in rbind(deparse.level, ...):
"number of columns of result, 6, is not a multiple of vector length 5 of arg 2"
A anova: 2 × 6
  Value        SE      Df Sum of Sq       F     Pr(>F)
<dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
Dosage1 -4.287576 1.837593 1.00 602.0182 5.444087 0.02065394
Residuals NA 195.000000 21563.49 NA NA NA

```

Hình 2.17: Phân tích t-test các nhóm

Với các giả định:

- H₀: Không có sự tương tác giữa 2 nhóm thuốc được nhắc đến.
- H₁: H₁: Có sự tương tác giữa 2 nhóm thuốc được nhắc đến

Với độ tin cậy 0.05, ta có nhận xét như sau:

- Nhóm A và S không có sự tương tác với nhau
- Nhóm A và T có sự tương tác với nhau
- Nhóm T và S có sự tương tác với nhau

- **Bước 3.2: Phân tích ảnh hưởng chính của Drug với hiệu quả của bài kiểm tra trí nhớ**

```

1 drug_model = aov(Diff ~ Drug, data = processed_islander)
2 summary(drug_model)

```

Kết quả:

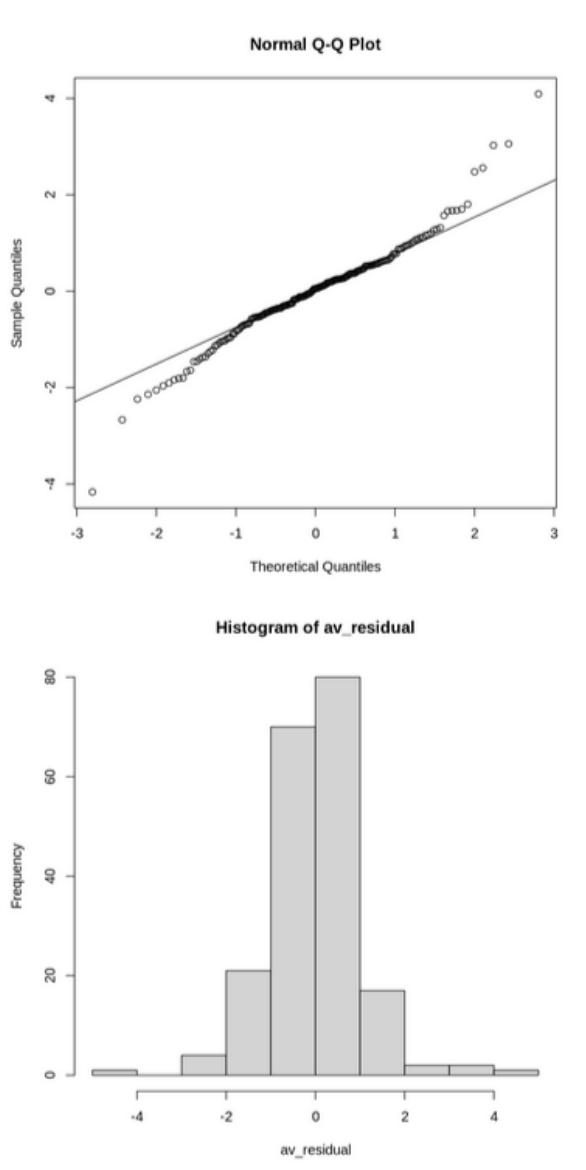
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
2	Drug	2	4305	2152.4	22.71	1.36e-09 ***
3	Residuals	195	18481	94.8		
4	---					
5	Signif. codes:	0	'***'	0.001	'**'	0.01
		'	*	0.05	.	0.1
		1				

Nhận xét: Với mức ý nghĩa 0.05, ta thấy rằng Dosage có ý nghĩa trong việc giải thích mô hình.

```
1 # Shapiro-Wilk test
2 av_residual = rstandard(drug_model)
3 shapiro.test(av_residual)
4
5 # Trực quan bằng QQ plot
6 qqnorm(av_residual)
7 qqline(av_residual)
8 hist(av_residual)
```

Kết quả:

```
1 Shapiro-Wilk normality test
2
3 data: av_residual
4 W = 0.96098, p-value = 2.777e-05
```



Hình 2.18: Đồ thị phần dư của ảnh hưởng chính giữa Drug và Diff

Với các giả định:

- H0: Phản dư tuân theo phân phối chuẩn.
- H1: Phản dư không tuân theo phân phối chuẩn

Nhận xét: với độ tin cậy 5% thì với giá trị p-value = 2.777e-05 chúng ta đủ cơ sở bác bỏ H0, vậy sai số có phân phối không chuẩn. Nhìn vào biểu đồ, ta thấy rằng ở phần đuôi kéo dài, có một vài điểm bị kéo lệch ra khỏi đường thẳng, Khả năng các điểm nhiễu chính là các điểm ngoại lệ (outliers), tuy nhiên, về mặt tổng quan, dữ liệu vẫn có dạng gần chuẩn.

```

1 # Kiểm định các nhóm có phương sai đồng nhất hay không
2 leveneTest(drug_model)

```

Kết quả:

```

1 A anova: 2 x 3
2   Df  F value Pr(>F)
3     <int>  <dbl>    <dbl>
4 group    2       19.06641      2.735522e-08
5          195       NA        NA

```

Với các giả định:

- H0: Các nhóm có phương sai đồng nhất
- H1: Các nhóm không có phương sai đồng nhất

Nhận xét: Với giá trị p-value = 2.735522e-08 < 0.05, ta đủ điều kiện bác bỏ H0, vậy các nhóm có phương sai không đồng nhất.

```

1 # Kiểm định tính độc lập của phần dư
2 durbinWatsonTest(drug_model)
3 plot(drug_model, 1)

```

Kết quả:

```

1 lag Autocorrelation D-W Statistic p-value
2   1           0.2969101       1.398674       0
3 Alternative hypothesis: rho != 0

```

Với các giả định:

- H0: Không có sự tương quan (độc lập)
- H1: Có sự tương quan (không độc lập)

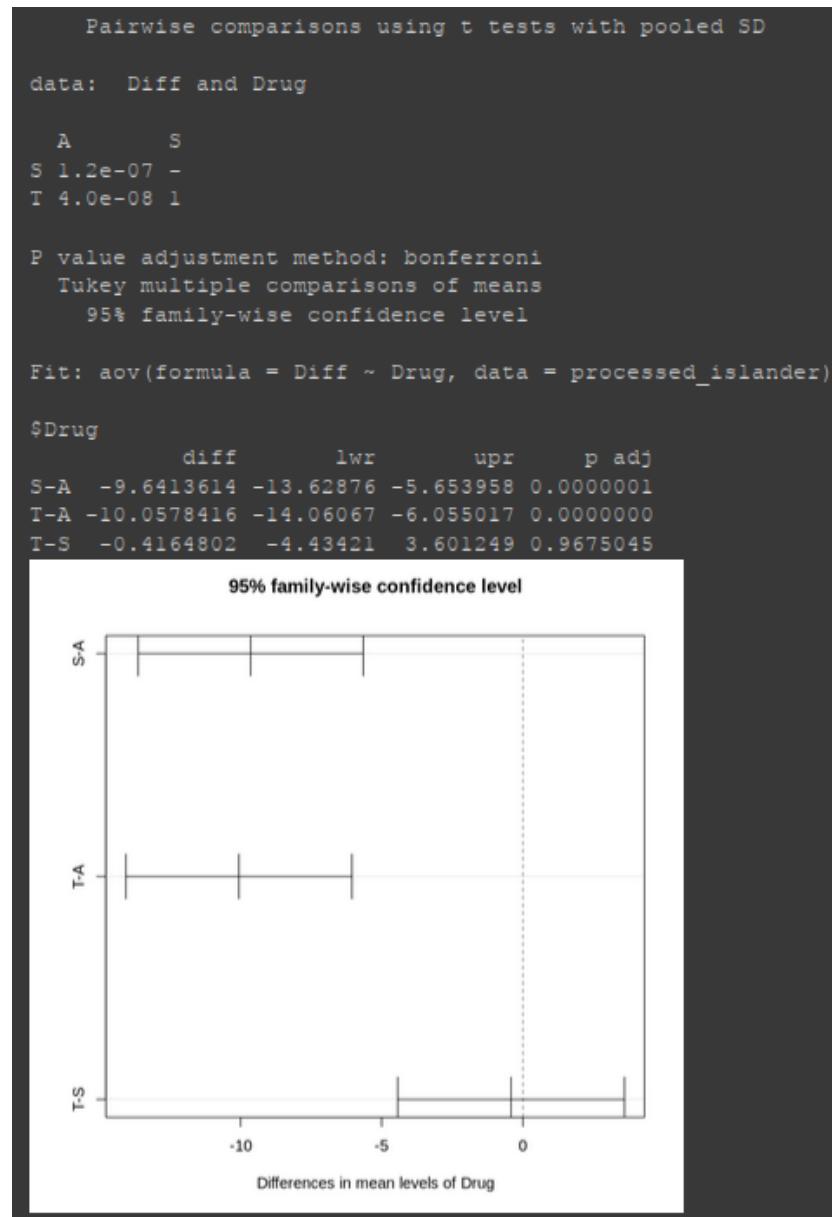
Nhận xét: Với giá trị p-value = 0 nên có sự tương quan dương. Mặc dù với điều kiện phương sai giữa các nhóm không đồng nhất nên sẽ không tiến hành phân tích ANOVA được, tuy nhiên về mặc trực quan hóa dữ liệu, ta thấy rằng đồ thị phân bố dạng gần chuẩn, nên ta sẽ tiếp tục đi phân tích các yếu tố ANOVA.

```

1 # Kiểm định độ hiệu quả trung bình giữa các nhóm thuốc
2 with(processed_islander, pairwise.t.test(Diff, Drug, p.adj
   = "bonferroni"))
3 TukeyHSD(aov(Diff~Drug, data=processed_islander), conf.
   level = 0.95)
4 plot(TukeyHSD(aov(Diff~Drug, data=processed_islander), conf
   .level = 0.95))

```

Kết quả:



Hình 2.19: Phân tích trung bình giữa các nhóm

Với các giả thuyết:

- H0: Các giá trị trung bình giữa các cặp bằng nhau
- H1: Các giá trị trung bình giữa các cặp không bằng nhau

Nhìn vào kết quả ta có:

- Nhóm T-S có p-value > 0.05 nên không đủ bác bỏ H0, vậy nhóm này có giá trị trung bình bằng nhau.
- Các nhóm còn lại p-value đều có giá trị nhỏ hơn 0.05 (độ tin cậy 95%) nên ta có cơ sở để bác bỏ H0. Vậy rõ ràng giữa các nhóm này có giá trị trung bình là khác nhau. Nghĩa là các nhóm thuộc khác nhau thì cho thấy mức độ ảnh hưởng đến bệnh nhân khác nhau. Nhìn vào kết quả và hình vẽ ta cũng thấy ngay giữa nhóm S-A và T-A có mức độ hiệu quả trung bình khác nhau, T-S có mức độ hiệu quả trung bình như nhau (đồ thị cắt điểm 0)

2.1.8. Xây dựng và kiểm định mô hình cộng (Additive model)

```
1 # Xây dựng mô hình cộng
2 add_model = lm(Diff ~ ., data=processed_islander)
3 add_model <- MASS::stepAIC(add_model, k = log(nrow(processed_
    islander)), trace = 0)
4 summary(add_model)
```

```
lm(formula = Diff ~ Dosage + Drug, data = processed_islander)

Residuals:
    Min      1Q  Median      3Q     Max 
-39.387 -4.702   0.161   5.341  36.099 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.9338     0.6720   4.366 2.06e-05 ***
Dosage1     -2.5514     0.9467  -2.695  0.00766 **  
Dosage2     -0.8414     0.9502  -0.885  0.37700    
Drug1       6.5744     0.9467   6.945 5.64e-11 ***
Drug2      -3.1051     0.9502  -3.268  0.00128 ** 
---
Signif. codes:  0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 '.' 1 

Residual standard error: 9.454 on 193 degrees of freedom
Multiple R-squared:  0.2429,    Adjusted R-squared:  0.2273 
F-statistic: 15.48 on 4 and 193 DF,  p-value: 5.296e-11
```

Hình 2.20: Kết quả mô tả của mô hình tuyến tính

Nhận xét: Với độ tin cậy 5%, các biến Dosage và Drug đều có ý nghĩa trong giải thích mô hình. Ta tiến hành kiểm định Shapiro và Breusch-Pagan

```

1 # Shapiro-Wilk test
2 av_residual = rstandard(add_model)
3 shapiro.test(av_residual)
4
5 # Trực quan bằng QQ plot
6 qqnorm(av_residual)
7 qqline(av_residual)
8 hist(av_residual)

```

Kết quả:

```

1 Shapiro-Wilk normality test
2
3 data: av_residual
4 W = 0.96409, p-value = 6.156e-05

```

Với các giả định:

- Phản dư H0: Tuân theo phân phối chuẩn
- H1: Phản dư Không tuân theo phân phối chuẩn

Nhận xét: với độ tin cậy 5% thì với giá trị p-value = 6.156e-05 chúng ta đủ cơ sở bác bỏ H0, vậy sai số có phân phối không chuẩn. Nhìn vào biểu đồ, ta thấy rằng ở phần đuôi kéo dài, có một vài điểm bị kéo lệch ra khỏi đường thẳng \rightarrow Khả năng các điểm nhiễu chính là các điểm ngoại lệ (outliners), tuy nhiên, về mặt tổng quan, dữ liệu vẫn có dạng gần chuẩn.

```

1 # Kiểm định tính độc lập của phản dư
2 durbinWatsonTest(add_model)
3 plot(add_model, 1)

```

Kết quả:

```

1 lag Autocorrelation D-W Statistic p-value
2    1          0.2550773      1.484015   0.002
3 Alternative hypothesis: rho != 0

```

Với các giả định:

- H0: Không có sự tương quan (độc lập)
- H1: Có sự tương quan (không độc lập)

Nhận xét: Với giá trị p-value = 0.02 < 0.05 nên có sự tương quan dương.

```

1 # Kiểm định Breusch-Pagan
2 bptest(add_model)

```

Kết quả:

```

1 studentized Breusch-Pagan test
2
3 data: add_model
4 BP = 9.0739, df = 4, p-value = 0.05928

```

Với các giả định:

- H0: phương sai không đổi
- H1: phương sai thay đổi

Nhận xét: Với p-value=0.059 > 0.05 thì ta không đủ điều kiện bác bỏ H0. Vậy phương sai của mô hình không thay đổi.

Như vậy, mô hình cộng được xây dựng như sau:

$$\text{Diff} = 2.9338 - 2.5514 \times \text{Dosage}_1 - 0.8414 \times \text{Dosage}_2 + 6.5744 \times \text{Drug}_1 - 3.1051 \times \text{Drug}_2$$

Với Adjusted R-squared = 0.2273, các biến giải thích được 22,73% ý nghĩa của mô hình, điều này có nghĩa rằng việc phục hồi trí nhớ sẽ bị chi phối bởi rất nhiều yếu tố (phần lớn là bản thân của người bệnh trầm cảm), chứ không chỉ mỗi tác động của thuốc. Sau đây sẽ là một số nhận xét của mô hình này:

- Hiệu ứng của liều lượng thứ nhất so với mức cơ bản. Giảm thời gian hoàn thành bài kiểm tra trí nhớ xuống 2.5514 giây, có ý nghĩa thống kê ($p < 0.01$).
- Dosage2: Hiệu ứng của liều lượng thứ hai so với mức cơ bản. Giảm thời gian hoàn thành bài kiểm tra trí nhớ xuống 0.8414 giây, nhưng không có ý nghĩa thống kê ($p = 0.377$).
- Drug1: Hiệu ứng của loại thuốc thứ nhất so với mức cơ bản. Tăng thời gian hoàn thành bài kiểm tra trí nhớ lên 6.5744 giây, có ý nghĩa thống kê cao ($p < 0.001$).
- Drug2: Hiệu ứng của loại thuốc thứ hai so với mức cơ bản. Giảm thời gian hoàn thành bài kiểm tra trí nhớ xuống 3.1051 giây, có ý nghĩa thống kê ($p < 0.01$).
- F-statistic: 15.48 với p-value = 5.296e-11, cho thấy mô hình tổng thể có ý nghĩa thống kê.

Kết luận: Nếu xem bản thân loại thuốc và liều thuốc tương tác một cách độc lập, thì sau đây là khuyến nghị cho bác sĩ:

- Liều lượng: Liều lượng thứ nhất có ảnh hưởng đáng kể đến thời gian hoàn thành bài kiểm tra trí nhớ, trong khi liều lượng thứ hai không có ảnh hưởng đáng kể. Khuyến nghị bác sĩ dùng liều lượng thứ nhất (liều lượng thấp).
- Loại thuốc: Cả hai loại thuốc đều có ảnh hưởng đáng kể đến thời gian hoàn thành bài kiểm tra trí nhớ, với loại thuốc thứ nhất tăng thời gian và loại thuốc thứ hai giảm thời gian. Khuyến nghị bác sĩ sử cho bệnh nhân sử dụng loại thuốc thứ hai (loại thuốc S)

Thực tế thì 2 nhân tố này ảnh hưởng trực tiếp đến nhau và cho kết quả khác với mô hình cộng (phân tích ảnh hưởng đơn ở phần trước) vì vậy, cần phải cẩn thận cân nhắc khi sử dụng thuốc tránh đem lại hậu quả không mong muốn ngoài tầm kiểm soát.

2.1.9. *Cải tiến mô hình*

Như chúng ta đã thấy ở các bước phân tích trên, khi phân tích ảnh hưởng chính cũng như xây dựng mô hình tuyến tính, có một số yêu cầu chưa thỏa mãn (ví dụ như tính chuẩn, tính độc lập của phương sai), vì vậy, trong phần này chúng ta sẽ tập trung xử lý dữ liệu cho phù hợp hơn. Như đã nhận định ở trên, hiện tại dữ liệu chúng ta đang tồn tại các điểm ngoại lai và cực ngoại lai, trong thí nghiệm này, chúng ta tiến hành loại bỏ các điểm này và tiến hành khảo sát. Ở phần này, tôi sẽ không trình bày chi tiết từng bước như trước (vì các bước thực hiện như nhau, mà thực thi đính kèm); chỉ trình bày những điểm thay đổi chính so với việc phân tích từ tập dữ liệu thô ban đầu.

Dầu tiên ta sẽ tiến hành khảo sát các điểm ngoại lai bằng lệnh sau

```

1 # Khảo sát ngoại lai theo biến diff
2 diff_data = processed_islander["Diff"]
3 outliers_index = list()
4 extreme_outliers_index = list()
5
6 for (i in 1:ncol(diff_data)) {
7   # Tính toán Q1, Q3 và IQR
8   Q1 = quantile(diff_data[, i], 0.25, na.rm = TRUE)
9   Q3 = quantile(diff_data[, i], 0.75, na.rm = TRUE)
10  IQR = Q3 - Q1
11
12  # Xác định ngoại lai
13  outliers_index_i = diff_data[, i] < (Q1 - 1.5 * IQR) | diff_
14    data[, i] > (Q3 + 1.5 * IQR)
15  # outliers_i = diff_data[diff_data[, i] < (Q1 - 1.5 * IQR) | 
16    diff_data[, i] > (Q3 + 1.5 * IQR), i]
17
18  outliers_index = c(outliers_index, outliers_index_i)
19  extreme_outliers_index = c(extreme_outliers_index, outliers_index_i)
20}
21
```

```

16 # Lưu trữ ngoại lai
17 field_name = names(diff_data)[i]
18 outliers_index[[field_name]] = which(outliers_index_i)
19
20 # Xác định cực ngoại lai
21 extreme_outliers_index_i = diff_data[, i] < (Q1 - 3 * IQR) |
22   diff_data[, i] > (Q3 + 3 * IQR)
23 extreme_outliers_index[[field_name]] = which(extreme_outliers_
24   _index_i)
25
26 # In kết quả theo từng biến ra màn hình
27 for (i in 1:ncol(diff_data)) {
28   print(paste("Biến:", names(diff_data)[i]))
29   print(paste("Số ngoại lai:", length(outliers_index[[names(
30     diff_data)[i]]])))
31   print(paste("Số cực ngoại lai:", length(extreme_outliers_
32     _index[[names(diff_data)[i]]])))
33 }
34
35 # Tìm tổng số quan trắc ngoại lai và cực ngoại lai thực sự
36 outliers = c()
37 extreme_outliners = c()
38 for (i in 1:ncol(diff_data)){
39   outliers = c(outliers, outliers_index[[names(diff_data)[i
40     ]]]))
41   extreme_outliners = c(extreme_outliners, extreme_outliers_
42     _index[[names(diff_data)[i]]])
43 }
44
45 outliers = unique(outliers)
46 extreme_outliners = unique(extreme_outliners)
47 print(paste("Tổng số ngoại lai:", length(outliers)))
48 print(paste("Tổng số cực ngoại lai:", length(extreme_outliners)
49   ))

```

Kết quả:

```
1 [1] "Biến: Diff"
2 [1] "Số ngoại lai: 20"
3 [1] "Số cực ngoại lai: 5"
4 [1] "Tổng số ngoại lai: 20"
5 [1] "Tổng số cực ngoại lai: 5"
```

Như vậy, tổng số ngoại lai và cực ngoại là là 25 samples (chiếm khoảng 12%). Ta tiến hành loại bỏ các điểm này

```
1 # Loại bỏ các điểm ngoại lai và cực ngoại lai
2 rm_outliner_islander = processed_islander[-extreme_outliners,]
3 rm_outliner_islander = rm_outliner_islander[-outliers,]
4
5 # Kiểm tra lại số lượng dữ liệu
6 dim(rm_outliner_islander)
7 str(rm_outliner_islander)
```

Kết quả

```
1 173 3
2 'data.frame':   173 obs. of  3 variables:
3 $ Dosage: Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
4   ...
5 $ Drug   : Factor w/ 3 levels "A","S","T": 1 1 1 1 1 1 1 1 1 1 ...
6   ...
7 $ Diff   : num  -2.3 -0.9 -4.6 -0.5 0.1 -8.3 11.9 -1.5 -11.2 12 ...
8   ...
```

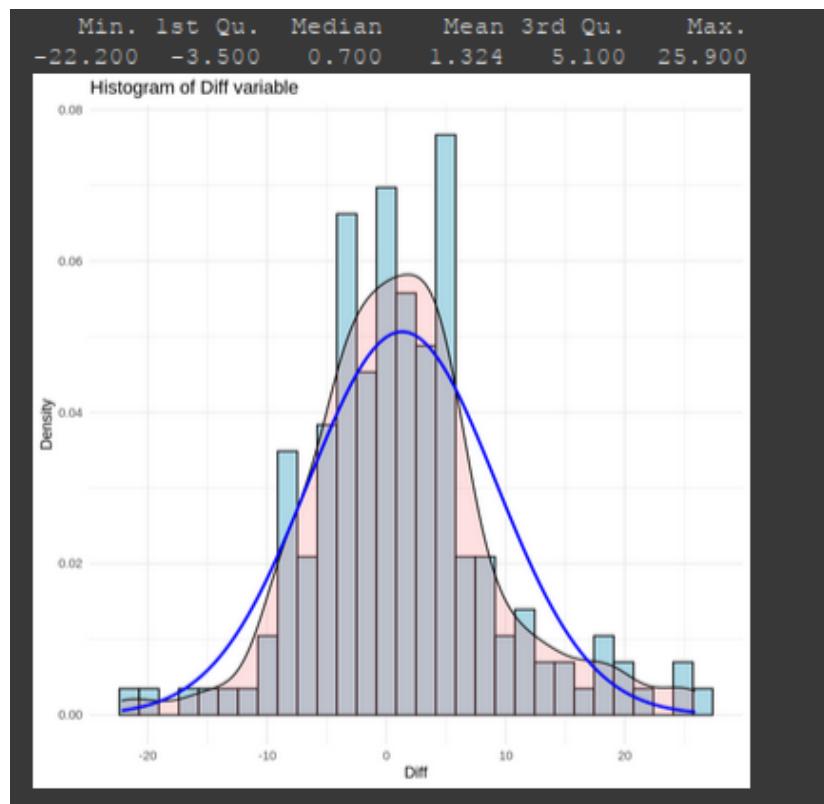
Như vậy, sau khi loại bỏ các điểm ngoại lai ta thu còn lại 173 samples. Ta tiến hành trực quan hóa đồ thị của dữ liệu

```

1 # Biến phụ thuộc Diff
2 ggplot(rm_outliner_islander, aes(x = Diff)) +
3   geom_histogram(aes(y = ..density..), bins = 30, color = "black",
4                 fill = "lightblue") +
5   geom_density(alpha = 0.2, fill = "#FF6666") +
6   stat_function(fun = dnorm, args = list(mean = mean(rm_
7     outliner_islander$Diff, na.rm = TRUE), sd = sd(rm_outliner
8     _islander$Diff, na.rm = TRUE)),
9                 color = "blue", size = 1) +
10  theme_minimal() +
11  labs(title = "Histogram of Diff variable", x = "Diff", y =
12        "Density")
13 summary(rm_outliner_islander$Diff)

```

Kết quả:



Hình 2.21: Trực quan hóa dữ liệu của biến Diff

Nhận xét: Sau khi loại bỏ các điểm ngoại lai và cực ngoại lai, ta thu được đồ thị gần chuẩn và có

hình dáng tốt hơn trước khi loại.

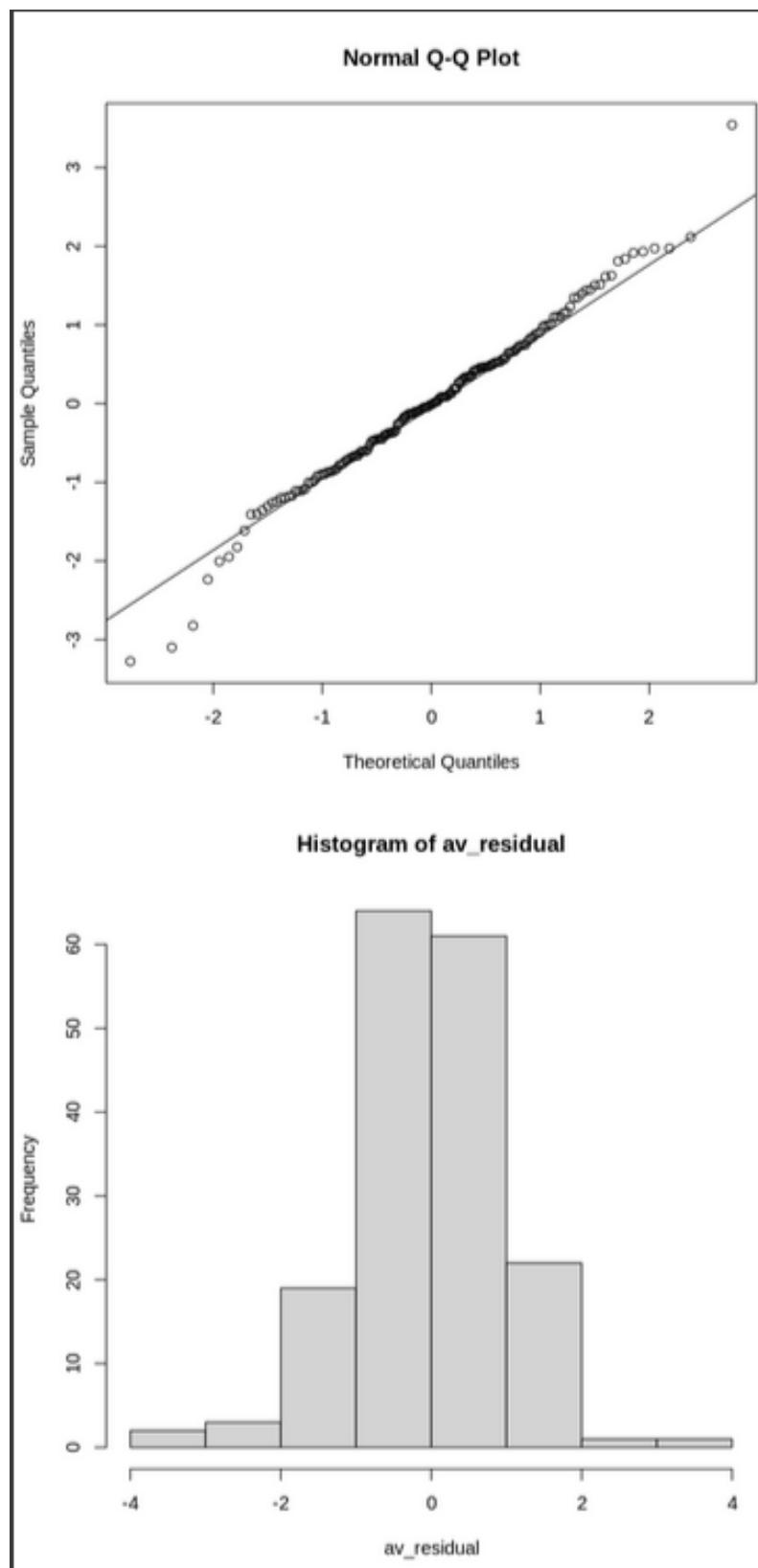
Tiếp theo chúng ta sẽ tiến hành xây dựng mô hình tương tác và kiểm định các giả thuyết

```
1 int_model = aov(Diff ~ Dosage * Drug, rm_outliner_islander)
```

```
1 # Kiểm định shapiro test
2 av_residual = rstandard(int_model)
3 shapiro.test(av_residual)
4
5 # Trực quan bằng QQ plot
6 qqnorm(av_residual)
7 qqline(av_residual)
8 hist(av_residual)
```

Kết quả:

```
1
2      Df Sum Sq Mean Sq F value    Pr(>F)
3 Dosage        2    190     94.8   2.115    0.124
4 Drug          2    957    478.4  10.670 4.40e-05 ***
5 Dosage:Drug   4   2166    541.6  12.079 1.25e-08 ***
6 Residuals    164   7354     44.8
7
8
9      Shapiro-Wilk normality test
10
11 data: av_residual
12 W = 0.98596, p-value = 0.08056
```



Hình 2.22: Phân phối của phần dư sau khi cải tiến

Với giả định

- H0: Phần dư tuân theo phân phối chuẩn
- H1: Phần dư không tuân theo phân phối chuẩn

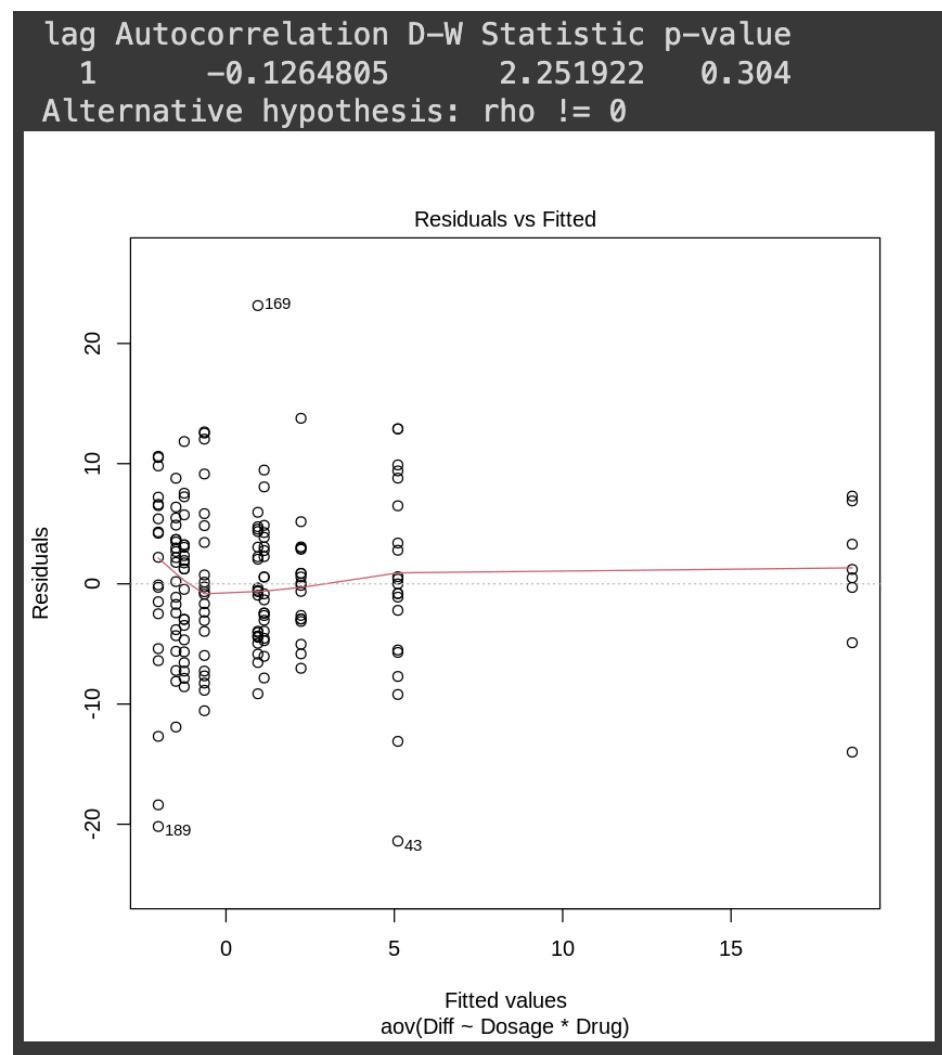
Với giả độ tin cậy 0.05 thì ta không đủ điều kiện bác bỏ H0, vậy phần dư tuân theo phân phối chuẩn. Mặc khác ta thấy rằng thấy bản thân liều lượng (Dosage) sẽ không tác động đến hiệu quả của người sử dụng thuốc, tuy nhiên chúng có mối liên hệ mật thiết (có tương tác) với loại thuốc.

Tiếp theo chúng ta đi kiểm định tính độc lập của phần dư:

```
1 # Kiểm định tính độc lập của phần dư
2 durbinWatsonTest(int_model)
3 plot(int_model, 1)
```

Kết quả

```
1 lag Autocorrelation D-W Statistic p-value
2     1      -0.1264805      2.251922    0.304
3 Alternative hypothesis: rho != 0
```



Hình 2.23: Đồ thị Residuals

Với mức ý nghĩa 5%, ta thấy rằng mô hình có phần dư độc lập. Tiếp tục Kiểm định các nhóm có phương sai đồng nhất hay không

```
1 # Kiểm định các nhóm có phương sai đồng nhất hay không  
2 leveneTest(int_model)
```

Kết quả

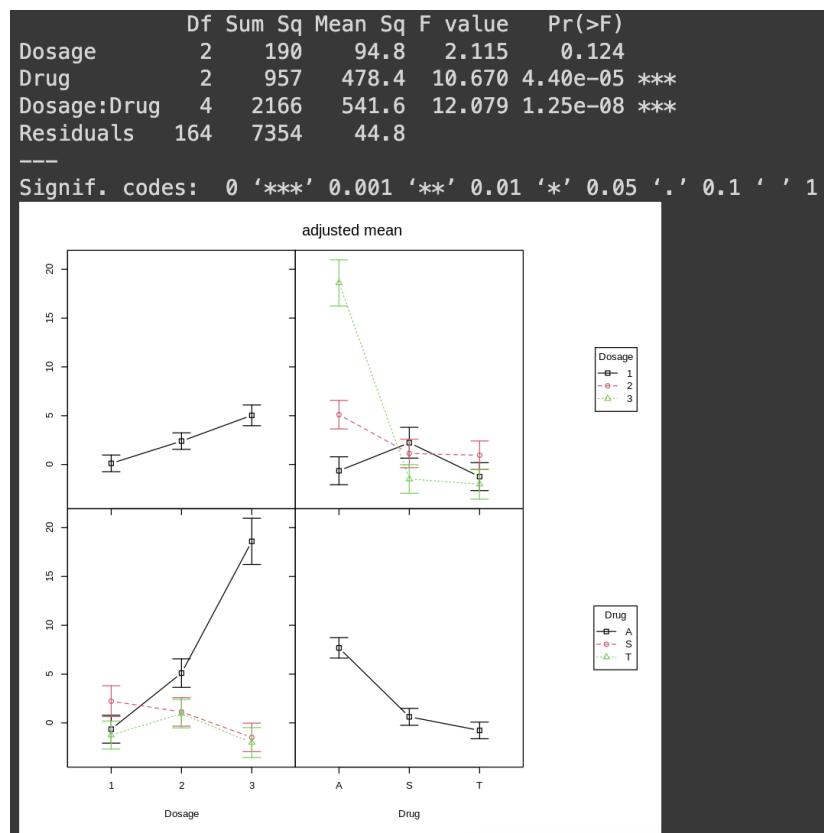
```
1 A anova: 2 x 3  
2 Df      F value Pr(>F)  
3 <int>    <dbl>   <dbl>  
4 group     8       1.440304      0.1833148  
5 164      NA      NA
```

Với mức ý nghĩa 5%, ta thấy mô hình có phương sai của các nhóm đồng nhất. Như vậy, ta đủ điều kiện để phân tích ANOVA. Bước tiếp theo, chúng ta sẽ tiến hành kiểm tra tương tác đơn và tương tác chính như phần trước.

- **Bước 1: Kiểm tra sự tương tác**

```
1 summary(int_model)  
2 plot(interactionMeans(int_model))
```

Kết quả:



Hình 2.24: Tương tác giữa Drug và Dosage

Nhận xét: Với mức ý nghĩa 5%, ta thấy giữa ‘Dosage’ và ‘Drug’ có sự tương tác với nhau ($p\text{-value}=1.25e-08$). Sự kết hợp giữa hàm lượng thuốc và loại thuốc có ảnh hưởng rất lớn đến thời gian hoàn thành bài kiểm tra, cho thấy rằng không chỉ từng yếu tố riêng lẻ mà sự kết hợp giữa chúng cũng rất quan trọng. Về phần nhận xét chi tiết xem lại phần đầu tiên vì kết quả biểu đồ giống với phân tích của phần đầu.

- **Bước 2: Phân tích ảnh hưởng đơn**

- a. **Phân tích ảnh hưởng đơn của liều lượng ở mỗi loại thuốc**

```
1 testInteractions(int_model, fixed = "Drug", across = "
  Dosage")
```

Kết quả:

"number of columns of result, 8, is not a multiple of vector length 6 of arg 2"								
A anova: 4 × 8								
	Dosage1	Dosage2	SE1	SE2	Df	Sum of Sq	F	Pr(>F)
A	-19.2409091	-13.495238	2.764648	2.782146	2	2176.58239	24.270232	1.754299e-09
S	3.7134921	2.619048	2.150909	2.066526	2	144.84001	1.615055	4.040531e-01
T	0.7696172	2.958145	2.097198	2.120211	2	95.99848	1.070442	4.040531e-01
Residuals	NA	NA	164.000000	7353.854655	NA	NA	NA	NA

Hình 2.25: Kết quả ảnh hưởng đơn của liều lượng ở mỗi loại thuốc

Giả định:

- * H0: Liều lượng Không ảnh hưởng đến hiệu quả thuốc
- * H1: Liều lượng Có ảnh hưởng đến hiệu quả của thuốc

Nhận xét: Với kết quả phân tích ta có một số nhận xét như sau, với độ tin cậy 5% thì:

- * Liều lượng có ảnh hưởng đến kết quả của loại thuốc A
- * Liều lượng không ảnh hưởng đến kết quả của loại thuốc S và T

b . Phân tích ảnh hưởng đơn của thuốc ở mỗi liều lượng

```
1 testInteractions(int_model, fixed = "Dosage", across =
  "Drug")
```

Kết quả:

Warning message in rbind(deparse.level, ...): "number of columns of result, 8, is not a multiple of vector length 6 of arg 2" A anova: 4 × 8								
	Drug1	Drug2	SE1	SE2	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	0.600000	3.4686869	2.019014	2.128227	2	132.2203	1.474337	2.319485e-01
2	4.157143	0.1857143	2.066526	2.066526	2	231.6200	2.582705	1.572813e-01
3	20.610526	0.5248120	2.822251	2.120211	2	2759.4956	30.770072	1.348841e-11
Residuals	NA	NA	164.000000	7353.854655	NA	NA	NA	NA

Hình 2.26: Kết quả ảnh hưởng đơn của thuốc ở mỗi liều lượng

Với giả định

- * H0: Các loại thuốc sẽ không tác động ở mỗi liều lượng
- * H1: Các loại thuốc sẽ có tác động ở mỗi liều lượng

Nhận xét: Với kết quả phân tích ta có một số nhận xét như sau, với độ tin cậy 5% thì: Hầu hết các loại thuốc sẽ có tác động ở liều lượng cao; liều lượng thấp và trung bình cho kết quả không đáng kể.

c. Phân tích ảnh hưởng đơn giữa các nhóm thuốc ứng với mỗi liều lượng

```
1 options(contrasts = c(unordered="contr.sum", ordered="contr.poly"))
2 A_vs_S = list(Drug = c(1, -1, 0))
3 A_vs_T = list(Drug = c(1, 0, -1))
4 S_vs_T = list(Drug = c(0, 1, -1))
5 # Nhóm A và S
6 testInteractions(int_model, custom = c(A_vs_S), fixed =
  "Dosage", adjustment = "bonferroni")
```

Anova. F test						
	Value	SE	Df	Sum of Sq	F	Pr (>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1 : Drug1	-2.868687	2.128227	1.000	81.47071	1.816897	5.386261e-01
2 : Drug1	3.971429	2.066526	1.000	165.60857	3.693275	1.690984e-01
3 : Drug1	20.085714	2.782146	1.000	2337.14601	52.121229	5.545574e-11
Residuals	NA	164.000000	7353.855	NA	NA	NA

Giả định:

- H0: Không có sự tác động giữa nhóm thuốc A và S ở các liều lượng
- H1: Có sự tác động giữa nhóm thuốc A và S ở các liều lượng

Nhận xét: Với độ tin cậy 5%

- Có sự tác động về hiệu quả khi sử dụng thuốc thuốc A và S ở các liều lượng cao.
- Ở liều lượng thấp và trung bình: Không có sự tác động

Hình 2.27: Ảnh hưởng giữa nhóm A và S

Với các giả định:

- * H0: Không có sự tác động giữa nhóm thuốc A và S ở các liều lượng
- * H1: Có sự tác động giữa nhóm thuốc A và S ở các liều lượng

Nhận xét: Với độ tin cậy 5% Có sự tác động về hiệu quả khi sử dụng thuốc thuốc A và S ở các liều lượng cao. Ở liều lượng thấp và trung bình: Không có sự tác động.

```

1 # Nhóm A và T
2 testInteractions(int_model, custom = c(A_vs_T), fixed =
    "Dosage", adjustment = "bonferroni")

```

Warning message in rbind(deparse.level, ...): "number of columns of result, 6, is not a multiple of vector length 5 of arg 2" A anova: 4 x 6						
	Value	SE	Df	Sum of Sq	F	Pr (>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1 : Drug1	0.600000	2.019014	1.000	3.9600	0.08831287	1.000000e+00
2 : Drug1	4.157143	2.066526	1.000	181.4593	4.04676517	1.376805e-01
3 : Drug1	20.610526	2.822251	1.000	2391.4317	53.33186783	3.478680e-11
Residuals	NA	164.000000	7353.855	NA	NA	NA

Hình 2.28: Ảnh hưởng giữa nhóm A và T

Với các giả định:

- * H0: Không có sự tác động giữa nhóm thuốc A và T ở các liều lượng
- * H1: Có sự tác động giữa nhóm thuốc A và T ở các liều lượng

Nhận xét: Với độ tin cậy 5% Có sự tác động về hiệu quả khi sử dụng thuốc thuốc A và T ở các liều lượng cao. Ở liều lượng thấp và trung bình: Không có sự tác động.

```

1 # Nhóm S và T
2 testInteractions(int_model, custom = c(S_vs_T), fixed =
    "Dosage", adjustment = "bonferroni")

```

<chèn ảnh>

Warning message in rbind(deparse.level, ...): "number of columns of result, 6, is not a multiple of vector length 5 of arg 2" A anova: 4 × 6						
	Value	SE	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1 : Drug1	3.4686869	2.128227	1.000	119.1147071	2.656404413	0.3151584
2 : Drug1	0.1857143	2.066526	1.000	0.3621429	0.008076231	1.0000000
3 : Drug1	0.5248120	2.120211	1.000	2.7473910	0.061270197	1.0000000
Residuals	NA	164.000000	7353.855	NA	NA	NA

Hình 2.29: Ảnh hưởng giữa S và T

Với các giả định:

- * H0: Không có sự tác động giữa nhóm thuốc S và T ở các liều lượng
- * H1: Có sự tác động giữa nhóm thuốc S và T ở các liều lượng

Nhận xét: Với độ tin cậy 5% thì không có sự khác biệt nào ở cả 3 liều lượng.

d. Phân tích ảnh hưởng đơn giữa các nhóm liều lượng ứng với mỗi loại thuốc

```

1 options(contrasts = c(unordered="contr.sum", ordered="
    contr.poly"))
2 low_vs_medium = list(Dosage = c(1, -1, 0))
3 low_vs_high = list(Dosage = c(1, 0, -1))
4 medium_vs_high = list(Dosage = c(0, 1, -1))

```

```

1 # Nhóm thấp và trung bình
2 testInteractions(int_model, custom = c(low_vs_medium),
    fixed = "Drug", adjustment = "bonferroni")

```

Kết quả:

Warning message in rbind(deparse.level, ...): "number of columns of result, 6, is not a multiple of vector length 5 of arg 2 A anova: 4 x 6						
	Value	SE	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
A : Dosage1	-5.745671	2.042908	1.000	354.69497	7.9101339	0.01654912
S : Dosage1	1.094444	2.150909	1.000	11.60953	0.2589068	1.00000000
T : Dosage1	-2.188528	2.042908	1.000	51.46095	1.1476424	0.85685163
Residuals	NA	164.000000	7353.855	NA	NA	NA

Hình 2.30: Ảnh hưởng giữa nhóm thấp và trung bình

Với các giả định:

- * H0: Không có sự tương tác nhau giữa liều thấp và liều trung bình
- * H1: Có sự tương tác nhau giữa liều thấp và liều trung bình

Nhận xét: Với độ tin cậy 5% Ở loại thuốc A: Có sự tương tác về hiệu quả khi sử dụng thuốc ở các liều lượng thấp và liều lượng trung bình; Ở loại thuốc S và T: Không có sự tương tác có ý nghĩa thống kê.

```
1 # Nhóm thấp và cao
2 testInteractions(int_model, custom = c(low_vs_high),
fixed = "Drug", adjustment = "none")
```

Kết quả:

Warning message in rbind(deparse.level, ...): "number of columns of result, 6, is not a multiple of vector length 5 of arg 2" A anova: 4 x 6						
	Value	SE	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
A : Dosage1	-19.2409091	2.764648	1.000	2171.91382	48.436348	7.777949e-11
S : Dosage1	3.7134921	2.150909	1.000	133.65715	2.980719	8.614496e-02
T : Dosage1	0.7696172	2.097198	1.000	6.03868	0.134670	7.141114e-01
Residuals	NA	164.000000	7353.855	NA	NA	NA

Hình 2.31: Ảnh hưởng giữa nhóm thấp và cao

Với các giả định:

- * H0: Không có tương tác nhau giữa liều thấp và liều cao
- * H1: Có tương tác nhau giữa liều thấp và liều cao

Nhận xét: Với độ tin cậy 5% Ở loại thuốc A: Có sự tương tác về hiệu quả khi sử dụng

thuốc ở các liều lượng thấp và liều lượng cao; Ở loại thuốc S và T: Không có sự tương tác có ý nghĩa thống kê.

```

1 # Nhóm trung bình và cao
2 testInteractions(int_model, custom = c(medium_vs_high),
                  fixed = "Drug", adjustment = "none")

```

Kết quả:

Warning message in rbind(deparse.level, ...): "number of columns of result, 6, is not a multiple of vector length 5 of arg 2"						
A anova: 4 x 6						
	Value	SE	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
A : Dosage1	-13.495238	2.782146	1.000	1055.04841	23.528877	2.841508e-06
S : Dosage1	2.619048	2.066526	1.000	72.02381	1.606220	2.068207e-01
T : Dosage1	2.958145	2.120211	1.000	87.28747	1.946618	1.648383e-01
Residuals	NA	164.000000	7353.855	NA	NA	NA

Hình 2.32: Tương tác giữa nhóm trung bình vào cao

Với các giả định:

- * H0: Không có sự tương tác giữa liều trung bình và liều cao
- * Có sự tương tác giữa liều trung bình và liều cao

Nhận xét: Với độ tin cậy 5% Ở loại thuốc A: Có sự tương tác về hiệu quả khi sử dụng thuốc ở các liều lượng trung bình và liều lượng cao; Ở loại thuốc S và T: Không có sự tương tác có ý nghĩa thống kê.

- **Kết luận:** Kết quả này giống với kết quả phân tích trước đó. Tuy nhiên các tính chất kiểm định về chuẩn cho đánh giá ANOVA đã cho kết quả tốt hơn so với trước khi chưa xử lý dữ liệu.

• Bước 3. Phân tích ảnh hưởng chính

- Phân tích ảnh hưởng chính của Dosage với hiệu quả của bài kiểm tra trí nhớ

```

1 dosage_model = aov(Diff~Dosage, data = rm_outliner_
                      islander)
2 summary(dosage_model)

```

Kết quả:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Dosage	2	190	94.84	1.539	0.218

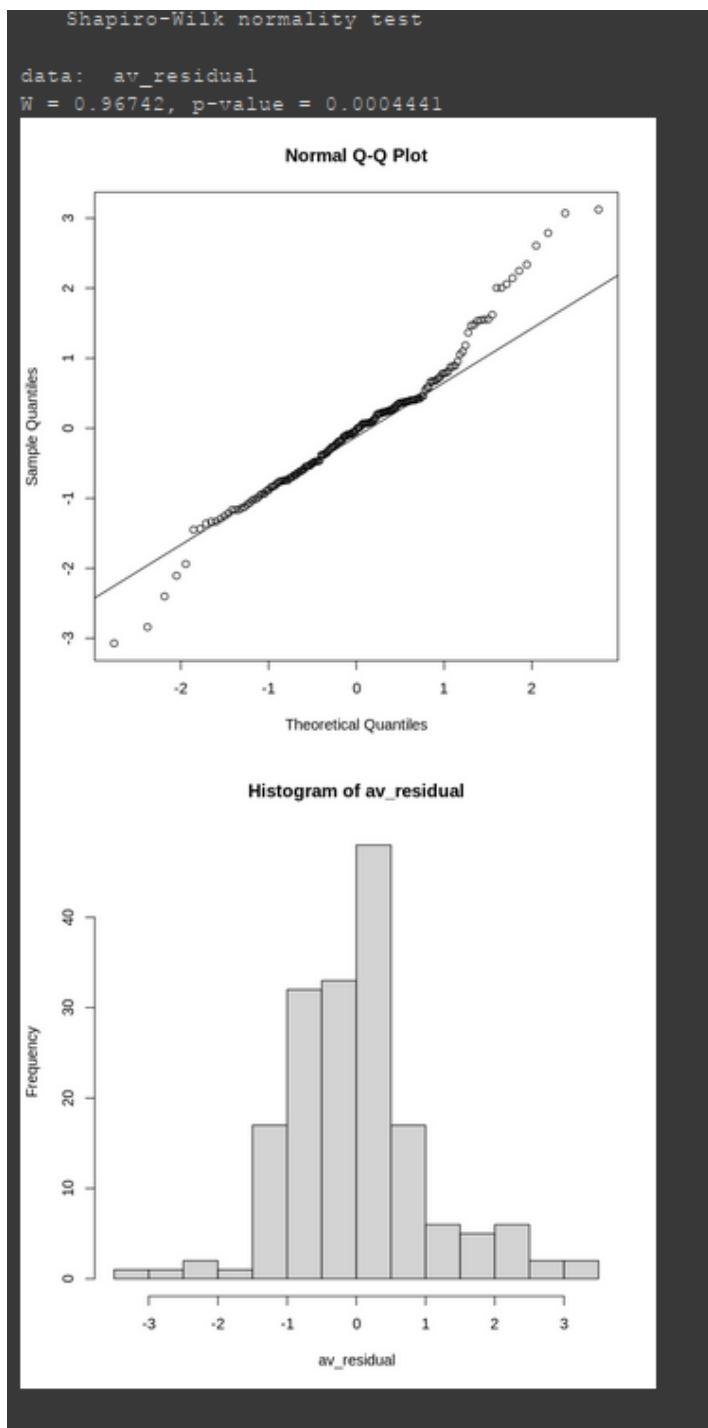
```
3 | Residuals     170   10477    61.63
```

Nhận xét: Với mức ý nghĩa 0.05, ta thấy rằng Dosage không có ý nghĩa trong việc giải thích mô hình. Theo nguyên tắc thì ta không cần phải đi kiểm định các giả thuyết cho biến này. Tuy nhiên chúng ta vẫn kiểm định để xem kết quả cải thiện như thế nào so với trước đó.

```
1 # Shapiro-Wilk test
2 av_residual = rstandard(dosage_model)
3 shapiro.test(av_residual)
4
5 # Trực quan bằng QQ plot
6 qqnorm(av_residual)
7 qqline(av_residual)
8 hist(av_residual)
```

Kết quả:

```
1      Shapiro-Wilk normality test
2 data: av_residual
3 W = 0.96742, p-value = 0.0004441
```



Hình 2.33: Shapiro-test

Nhận xét: Giá trị p-value đã tăng lên rất nhiều (mặc dù < 0.05), hình dáng đồ thị gần chuẩn hơn so với trước khi chưa xử lý dữ liệu.

```
1 # Kiểm định các nhóm có phương sai đồng nhất hay không
2 leveneTest(dosage_model)
```

Kết quả:

A anova:			
	Df	F value	Pr(>F)
	<int>	<dbl>	<dbl>
group	2	4.442547	0.01316301
	170	NA	NA

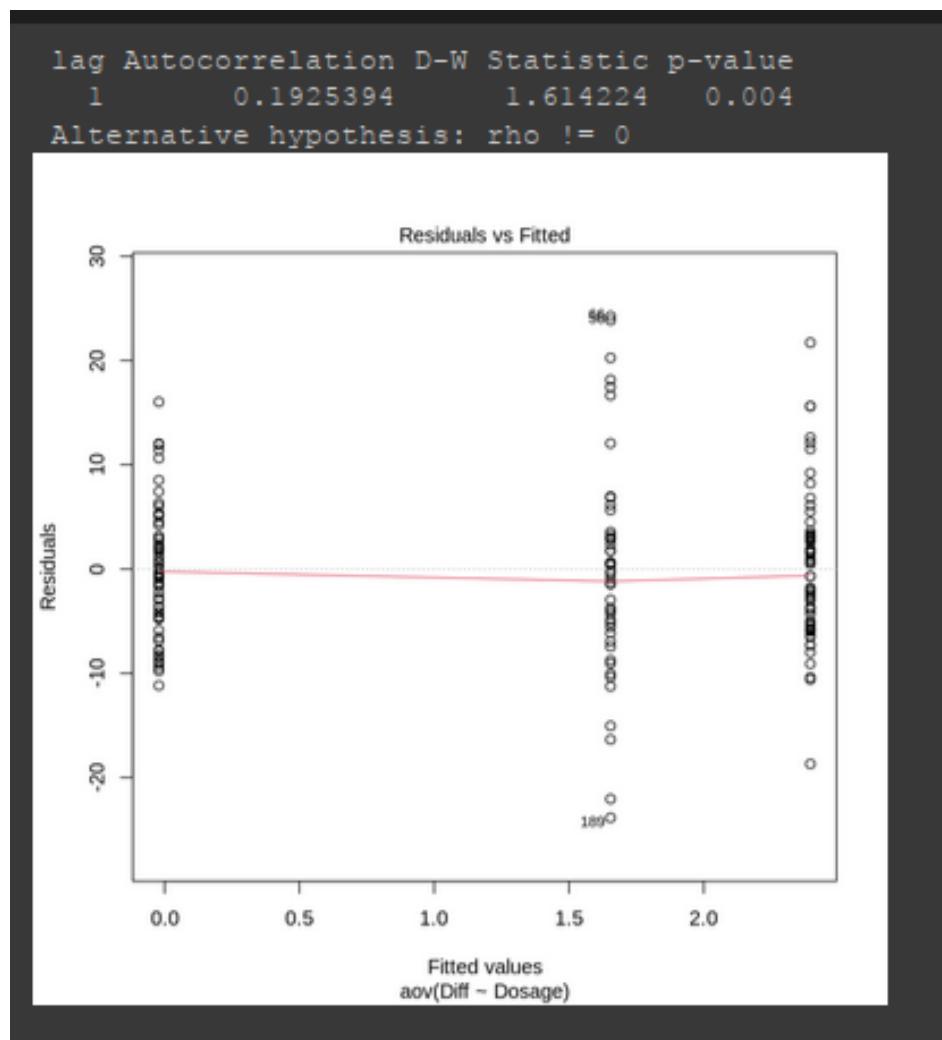
Nhận xét: Với các giả định:

- Các nhóm có phương sai đồng nhất
- Các nhóm không có phương sai đồng nhất

Nhận xét: Với giá trị p-value = 0.013 > 0.05, ta không điều kiện bác bỏ H0, vậy các nhóm có phương sai đồng nhất (trước đó là không đồng nhất) trước đó điều kiện này không thỏa mãn.

```
1 # Kiểm định tính độc lập của phần dư
2 durbinWatsonTest(dosage_model)
3 plot(dosage_model, 1)
```

Kết quả:



Hình 2.34: Kiểm định độc lập phần dư

Nhận xét: Với các giả định:

- H0: Không có sự tương quan (độc lập)
- H1: Có sự tương quan (không độc lập)

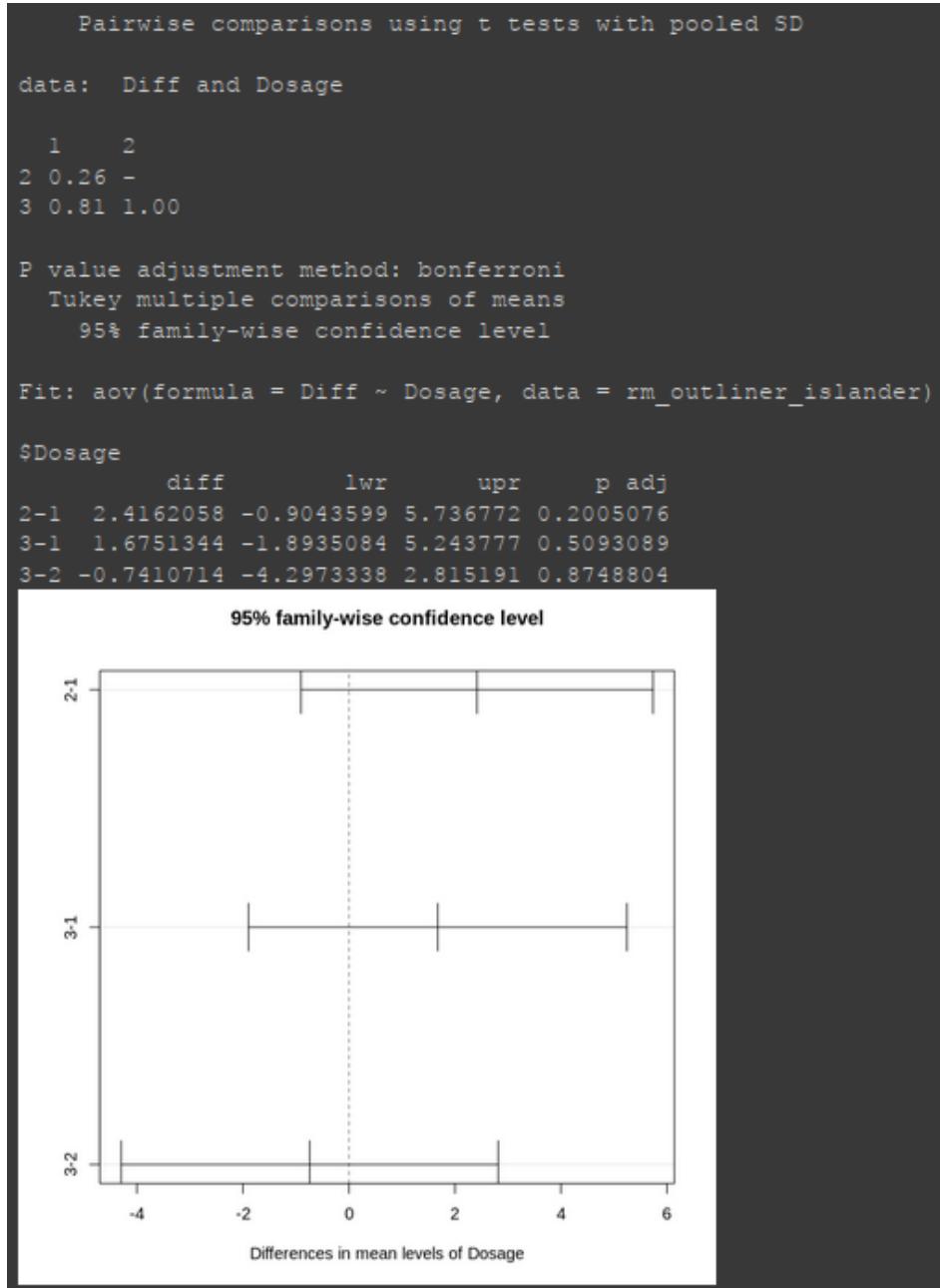
Nhận xét: VỚI giá trị p-value = 0.02 nên có sự tương quan dương, tuy nhiên kết quả này lớn hơn kết quả trước đó (=0).

```

1 # Kiểm định trung bình giữa các nhóm liều lượng
2 with(rm_outliner_islander, pairwise.t.test(Diff, Dosage, p.
  adj = "bonferroni"))
3 TukeyHSD(aov(Diff~Dosage, data=rm_outliner_islander), conf.
  level = 0.95)
4 plot(TukeyHSD(aov(Diff~Dosage, data=rm_outliner_islander),
  conf.level = 0.95))

```

Kết quả:



Hình 2.35: Kiểm định trung bình

Với các giả định:

- H0: Các giá trị trung bình giữa các cặp bằng nhau
- H1: Các giá trị trung bình giữa các cặp không bằng nhau

Nhận xét:Các cặp có p-value đều có giá trị lớn hơn 0.05 (độ tin cậy 95%) nên ta cơ sở để bác bỏ H0. Vậy rõ ràng giữa các nhóm này có giá trị trung bình là như nhau. Để rõ hơn, ta tiến hành kiểm định Tukey's. Nhìn vào kết quả và hình vẽ ta cũng thấy ngay mức độ hiệu quả trung bình như nhau ở 3 nhóm (đồ thị cát điểm 0). **Kết quả trước đó cho ta thấy rằng 3-2 và 1-2 có mức độ hiệu quả trung bình như nhau và 3-1 là khác nhau.**

```
1 # ttest
2 A_vs_S = list(Dosage = c(1, -1, 0))
3 A_vs_T = list(Dosage = c(1, 0, -1))
4 S_vs_T = list(Dosage = c(0, 1, -1))
5 testInteractions(dosage_model, custom = A_vs_S, adjustment
6   = 'bonferroni')
7 print("-----")
8 testInteractions(dosage_model, custom = A_vs_T, adjustment
9   = 'bonferroni')
10 print("-----")
11 testInteractions(dosage_model, custom = S_vs_T, adjustment
12   = 'bonferroni')
```

```

Warning message in rbind(deparse.level, ...):
"number of columns of result, 6, is not a multiple of vector length 5 of arg 2"
A anova: 2 × 6
  Value        SE      Df Sum of Sq      F Pr(>F)
<dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
Dosage1 -2.416206 1.404387 1.00 182.4274 2.960017 0.08716664
Residuals NA 170.000000 10477.19 NA NA NA
[1] -----
Warning message in rbind(deparse.level, ...):
"number of columns of result, 6, is not a multiple of vector length 5 of arg 2"
A anova: 2 × 6
  Value        SE      Df Sum of Sq      F Pr(>F)
<dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
Dosage1 -1.675134 1.509308 1.00 75.91709 1.23181 0.2686228
Residuals NA 170.000000 10477.19 NA NA NA
[1] -----
Warning message in rbind(deparse.level, ...):
"number of columns of result, 6, is not a multiple of vector length 5 of arg 2"
A anova: 2 × 6
  Value        SE      Df Sum of Sq      F Pr(>F)
<dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
Dosage1 0.7410714 1.504072 1.00 14.96163 0.2427633 0.6228532
Residuals NA 170.000000 10477.19 NA NA NA

```

Hình 2.36: Kết quả t-test

Với các giả định:

- H0: Không có sự tương tác giữa 2 nhóm thuốc được nhắc đến
- H1: Có sự tương tác giữa 2 nhóm thuốc được nhắc đến

Nhận xét: Với p-value=0.05, ta có kết luận như sau: cả 3 nhóm đều có sự tương tác mang ý nghĩa thống kê (giống kết quả trước đó).

- Phân tích ảnh hưởng chính của Drug với hiệu quả của bài kiểm tra trí nhớ

```

1 drug_model = aov(Diff~Drug, data = rm_outliner_islander)
2 summary(drug_model)

```

Kết quả:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
2 Drug	2	896	447.8	7.791	0.000579 ***
3 Residuals	170	9771	57.5		
4 ---					
5 Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 '
	1				

Nhận xét: Với mức ý nghĩa 0.05, ta thấy rằng Drug có ý nghĩa trong việc giải thích mô hình.

```

1 # Shapiro-Wilk test
2 av_residual = rstandard(drug_model)
3 shapiro.test(av_residual)
4
5 # Trực quan bằng QQ plot
6 qqnorm(av_residual)
7 qqline(av_residual)
8 hist(av_residual)

```

Kết quả:

```

1 Shapiro-Wilk normality test
2
3 data: av_residual
4 W = 0.9859, p-value = 0.07921

```

Với các giả định:

- H0: Tuân theo phân phối chuẩn
- H1: Không tuân theo phân phối chuẩn

Nhận xét: với độ tin cậy 5% thì với giá trị p-value = 0.07921 chúng ta không đủ cơ sở bác bỏ H0, vậy sai số có phân phối chuẩn. Nhìn vào biểu đồ, ta thấy rằng ở phần đuôi kéo dài, có một vài điểm bị kéo lệch ra khỏi đường thẳng về mặt tổng quan, dữ liệu vẫn có dạng gần chuẩn (trước đó là không chuẩn)

```

1 # Kiểm định các nhóm có phương sai đồng nhất hay không
2 leveneTest(drug_model)

```

Kết quả:

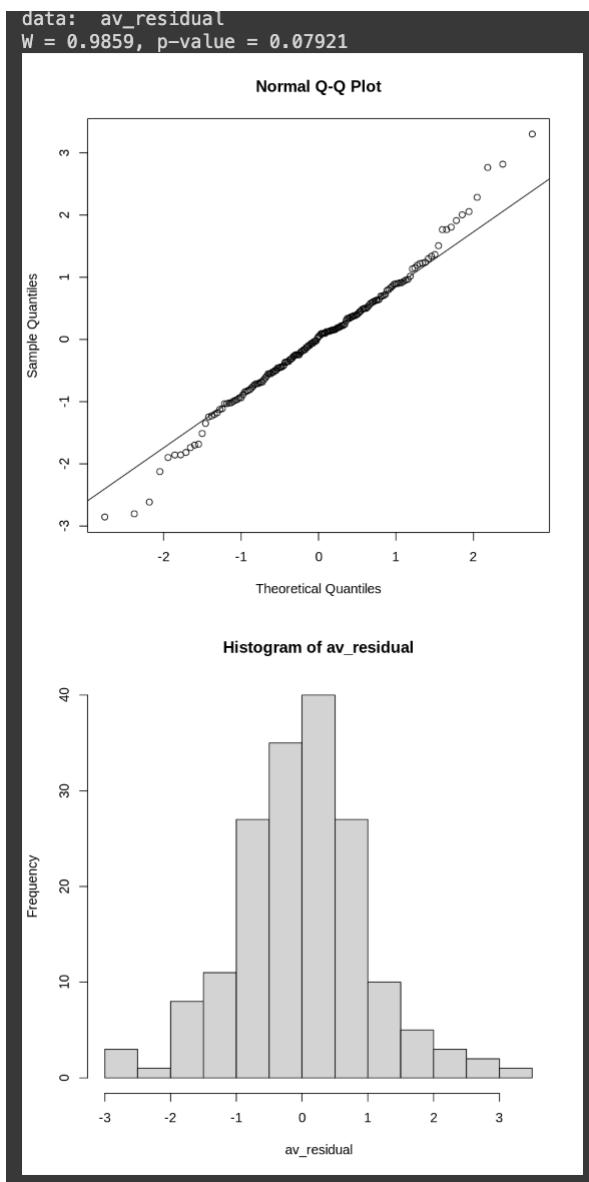
```

1 A anova:
2 2 x 3      Df      F value Pr(>F)
3          <int>    <dbl>    <dbl>
4 group      2       11.12926     2.87001e-05
5          170      NA        NA

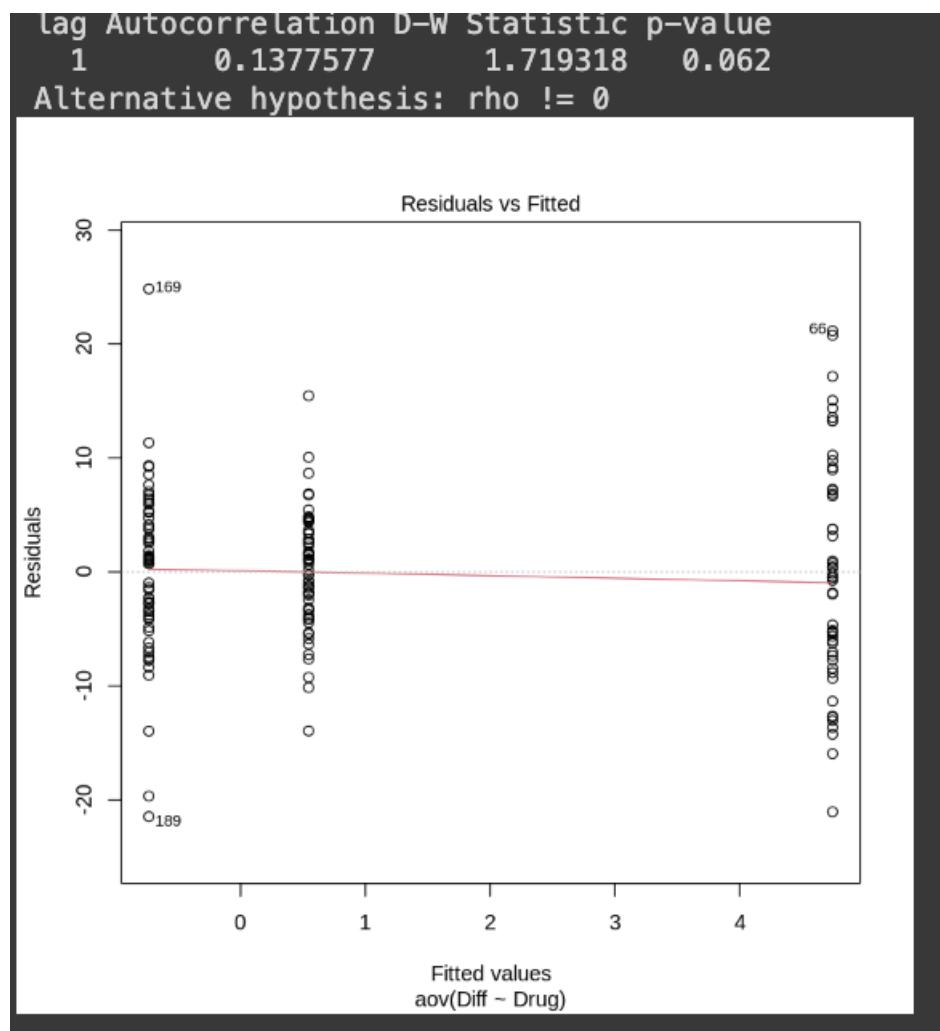
```

Với các giả định:

- H0: Các nhóm có phương sai đồng nhất
- H1: Các nhóm không có phương sai đồng nhất



Hình 2.37: Shapiro-Wilk test và đồ thị phân dữ



Hình 2.38: Kiểm định tính độc lập của phần dư

Nhận xét: Nhận xét: Với giá trị p-value = 2.87001e-05 < 0.05, ta đủ điều kiện bác bỏ H₀, vậy các nhóm có phương sai không đồng nhất, tuy nhiên kết quả có giá trị p-value cao hơn trước (2.735522e-08).

```
1 # Kiểm định tính độc lập của phần dư
2 durbinWatsonTest(drug_model)
3 plot(drug_model, 1)
```

Với các giả định:

- H₀: Không có sự tương quan (độc lập)
- H₁: Có sự tương quan (không độc lập)

Nhận xét: Với giá trị p-value = 0.062 nên không có sự tương quan (trước đó là tương quan dương).

```

1 # Kiểm định độ hiệu quả trung bình giữa các nhóm thuốc
2 with(rm_outliner_islander, pairwise.t.test(Diff, Drug, p.
3   adj = "bonferroni"))
4 TukeyHSD(aov(Diff~Drug, data=rm_outliner_islander), conf.
5   level = 0.95)
6 plot(TukeyHSD(aov(Diff~Drug, data=rm_outliner_islander),
7   conf.level = 0.95))

```

Với các giả định:

- H0: Các giá trị trung bình giữa các cặp bằng nhau
- H1: Các giá trị trung bình giữa các cặp không bằng nhau

Nhận xét:

- Nhìn vào kết quả ta có: Nhóm T-S có p-value > 0.05 nên không đủ bác bỏ H0, vậy nhóm này có giá trị trung bình bằng nhau; Các nhóm còn lại p-value đều có giá trị nhỏ hơn 0.05 (độ tin cậy 95%) nên ta có cơ sở để bác bỏ H0. Vậy rõ ràng giữa các nhóm này có giá trị trung bình là khác nhau.
- Nhìn vào kết quả và hình vẽ ta cũng thấy ngay giữa nhóm S-A và T-A có mức độ hiệu quả trung bình khác nhau, T-S có mức độ hiệu quả trung bình như nhau (đồ thị cắt điểm 0) (giống kết quả phân tích trước đó).
- **Kết luận:** các tính chất kiểm định về chuẩn cho đánh giá ANOVA và kiểm định ảnh hưởng chính đã cho kết quả tốt hơn so với trước khi chưa xử lý dữ liệu.

- **Bước 4: Xây dựng và kiểm định mô hình cộng (Additive model)**

```

1 add_model = lm(Diff~, data=rm_outliner_islander)
2 add_model <- MASS::stepAIC(add_model, k = log(nrow(rm_
3   outliner_islander)), trace = 0)
4 summary(add_model)
5 add_model$coefficients

```

Kết quả:

```

1
2 Call:
3 lm(formula = Diff ~ Drug, data = rm_outliner_islander)
4
5 Residuals:
6      Min       1Q     Median       3Q      Max

```

```

7 -21.4645   -4.4450    0.3569     4.3550   24.8355
8
9 Coefficients:
10                         Estimate Std. Error t value Pr(>|t|)
11 (Intercept)            1.5176    0.5785   2.623 0.009502 ** 
12 Drug1                  3.2256    0.8428   3.827 0.000182 *** 
13 Drug2                 -0.9726    0.8087  -1.203 0.230799
14 ---
15 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
16
17 Residual standard error: 7.581 on 170 degrees of freedom
18 Multiple R-squared:  0.08396,    Adjusted R-squared:
19                           0.07318
20
21 (Intercept)
22           1.51755112797807
23 Drug1
24           3.22558612692389
25 Drug2
26          -0.972551127978073

```

Nhận xét: Với p-value=5%, chỉ có biến Drug có ý nghĩa trong việc giải thích mô hình. như vậy việc kiểm định mô hình cộng giống như phân tích ảnh hưởng chính của biến Drug

Như vậy, mô hình cộng được xây dựng như sau:

$$\text{Diff} = 1.517 + 3.225 \times \text{Drug1} - 0.972 \times \text{Drug2}$$

- Drug1: Hệ số cho Drug1 là 3.2256. Điều này cho thấy rằng khi sử dụng loại thuốc thứ nhất, sự khác biệt trong thời gian hoàn thành bài kiểm tra tăng thêm 3.2256 giây so với không sử dụng thuốc. Hệ số này có ý nghĩa thống kê ($p\text{-value} = 0.000182 < 0.001$) - Drug2:Hệ số cho Drug2 là -0.9726. Điều này cho thấy rằng khi sử dụng loại thuốc thứ hai, sự khác biệt trong thời gian hoàn thành bài kiểm tra giảm đi 0.9726 giây so với không sử dụng thuốc. Tuy nhiên, hệ số này không có ý nghĩa thống kê ($p\text{-value} = 0.230799 > 0.05$). - Mô hình tổng thể có ý nghĩa: F-statistic cho thấy mô hình tổng thể có ý nghĩa thống kê, tuy nhiên, Multiple R-squared thấp cho thấy mô hình chỉ giải thích được một phần nhỏ sự biến thiên của Diff

Kết luận: Nếu xem bản thân loại thuốc và liều thuốc tương tác một cách độc lập, thì sau đây là khuyến nghị cho bác sĩ: Nên sử dụng loại thuốc 2 (thuốc S) cho bệnh nhân. Thực tế thì việc sử

dụng thuốc cần đánh giá ở nhiều khía cạnh (ví dụ như phân tích ảnh hưởng đơn cho thấy tương tác mạnh với liều lượng) Vì vậy, cần phải cẩn thận cân nhắc khi sử dụng thuốc tránh đem lại hậu quả không mong muốn ngoài tầm kiểm soát.

Như vậy về tổng thể sao khi loại bỏ các điểm ngoại lai và cực ngoại lai, về vieeck thống kê và phân tích ANOVA đã cho ra một mô hình có các yếu tố thỏa mãn các yếu tố kiểm định về chuẩn hơn, trong TH không chuẩn nhưng chỉ số so với trước là tốt hơn

2.2. Phân tích phim truyền thông và xã hội

2.2.1. Giới thiệu chung

Trong những năm gần đây, các nhà phân tích và nhà đầu tư ngày càng quan tâm đến việc đánh giá rủi ro tài chính trong sản xuất phim. Nghiên cứu này sử dụng phân tích hồi quy tuyến tính bội để dự đoán thành công về mặt tài chính của phim và nghiên cứu mối quan hệ giữa số lần chiếu và năm.

2.2.2. Phát biểu bài toán

Mục tiêu chính của phần này là khám phá và phân tích tổng doanh thu của phim trong hai năm 2014 và 2015 cũng như kiểm tra mối quan hệ và ý nghĩa của một số biến giải thích. Hơn nữa, đồ án xây dựng một mô hình hồi quy tối ưu để đưa ra dự đoán về sự thành công về mặt tài chính, tức là tổng doanh thu của một bộ phim trong hai năm 2014 và 2015.

2.2.3. Giới thiệu về dữ liệu

Bộ dữ liệu phim truyền thông và xã hội (conventional and social media dataset) được sử dụng trong đồ án này có cấu trúc tương đối đơn giản mà một số người có kiến thức về phim truyền hình cũng có thể hiểu được. Vấn đề chính của bộ dữ liệu là missing values, và chúng tôi sẽ cố gắng xử lý nó bằng một số kỹ thuật đã biết.

Ngành công nghiệp điện ảnh là một ngành đóng góp đáng kể cho nền kinh tế của một quốc gia và là một nhà tuyển dụng lớn tại Hoa Kỳ. Do chi phí lớn liên quan đến sản xuất phim, các nhà phân tích cần nghiên cứu và hiểu các biến số chính góp phần vào thành công về mặt thương mại và tài chính của một bộ phim. Đồ án có thể cung cấp thông tin chi tiết về các tính năng chính góp phần vào thành công về mặt tài chính của các bộ phim và thúc đẩy nghiên cứu trong tương lai để xem xét mối quan hệ giữa các biến giải thích đặc biệt độc đáo trong tập dữ liệu. Hơn nữa nó còn có thể giúp các nhà sản xuất phim xác định những tính năng nào cần tập trung vào trong giai đoạn quảng bá để cải thiện thành công của bộ phim.

2.2.4. Khám phá và tiền xử lý dữ liệu

Đọc dữ liệu và một số kiểm tra khởi đầu

Trước hết, ta đọc dữ liệu

```
1 # Đọc dữ liệu từ tập tin
2 raw_data = read_excel("../data/part1/CSM.xlsx", sheet = 1)
3 str(raw_data)
4
5 # Thay đổi tên biến `Aggregate Followers` thành `
6 AggregateFollowers
6 names(raw_data)[names(raw_data) == 'Aggregate Followers'] <- '
7 AggregateFollowers'
```

Dữ liệu này không có hiện tượng trùng lặp. Dựa trên thông tin của tập dữ liệu, ta thấy mỗi dòng mang ý nghĩa khác nhau, tức là mỗi quan trắc độc lập nhau. Ý nghĩa từng cột như sau:

- **Movie:** tên phim
- **Year:** năm phát hành
- **Ratings:** điểm đánh giá
- **Genre:** thể loại phim
- **Gross:** tổng doanh thu
- **Budget:** tổng chi phí
- **Screens:** số rạp chiếu
- **Sequel:** phần phim
- **Sentiment:** ý kiến khán giả
- **Views:** số lượt xem
- **Likes:** số lượt thích
- **Dislikes:** số lượt chê
- **Comments:** số bình luận
- **Aggregate Followers:** số người theo dõi

Ta thấy biến Dislikes thể hiện ý nghĩa tương tự biến Likes nhưng có chiều hướng ngược lại. Nên ta có thể loại bỏ biến này khỏi tập dữ liệu.

Các cột với kiểu dữ liệu số phân bố như thế nào?

Ta kiểm tra một số thông tin thống kê mô tả của bộ dữ liệu

A tibble: 12 × 7

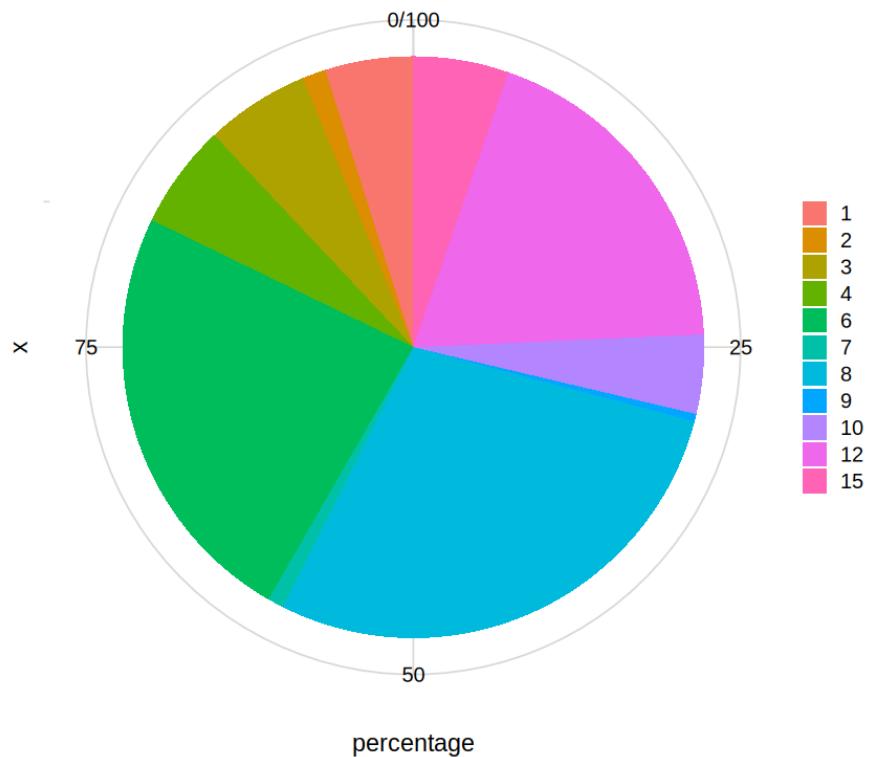
variable	missing	min	lower	median	upper	max
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1 Year	0	2014	2014	2014	2015	2.02e3
2 Ratings	0	3.1	5.8	6.5	7.1	8.7 e0
3 Gross	0	2470	10300000	37400000	89350000	6.43e8
4 Budget	0.4	70000	9000000	28000000	65000000	2.5 e8
5 Screens	4.3	2	449	2777	3372	4.32e3
6 Sequel	0	1	1	1	1	7 e0
7 Sentiment	0	-38	0	0	5.5	2.9 e1
8 Views	0	698	623302	2409338	5217380.	3.26e7
9 Likes	0	1	1776.	6096	15248.	3.71e5
10 Dislikes	0	0	106.	341	698.	1.40e4
11 Comments	0	0	248.	837	2137	3.84e4
12 AggregateFollowers	15.2	1066	183025	1052600	3694500	3.10e7

Nhận xét:

- Có hiện tượng missing values đối với cột AggregateFollowers, Screens và Budget. Cụ thể, ta thấy biến Aggregate Followers có tỷ lệ missing 15.2%, biến Screens có tỷ lệ 4.3% và biến Budget có tỷ lệ missing 0.4%.
- Có những bộ phim không có likes/ dislikes/ comments, ta sẽ loại bỏ những dòng này.

Các cột với kiểu dữ liệu phân loại phân bố như thế nào?

Khảo sát cột Genre thể hiện các thể loại phim, ta trực quan bằng biểu đồ tròn dưới đây.



Hình 2.39: Tỷ lệ các thể loại phim.

Nhận xét:

- Các thể loại phim phân bố không đều nhau
-

Xử lý dữ liệu bị thiếu

Trong đồ án này, chúng tôi khảo sát nhiều phương pháp xử lý dữ liệu bị thiếu như sau:

- Chèn không (zeros imputed)
- Chèn trung bình (means imputed)
- Chèn trung vị (median imputed)
- Điền các giá trị bị thiếu dựa trên PCA (imputed PCA)

Dựa trên kết quả thực nghiệm, chúng tôi chọn điền các giá trị bị thiếu dựa trên PCA vì có kết quả xây dựng mô hình tốt. Chi tiết các kỹ thuật khác được trình bày trong các file code.

Các bước thực hiện điền các giá trị bị thiếu dựa trên PCA:

- Bước 1: ước lượng số thành phần chính

```

1 # Ước lượng thành phần chính
2 nPCs <- estim_ncpPCA(raw_data[, -c(1)])
3 print(nPCs)

```

- Bước 2: điền các giá trị bị thiếu

```

1 # Xử lý missing value
2 processed_data <- imputePCA(raw_data[, -c(1)], ncp = nPCs$ncp,
3                               scale = TRUE)
4 processed_data <- processed_data$completeObs

```

Dánh giá kết quả bằng cách trực quan hóa phân phối trước và sau khi thực hiện việc điền các giá trị bị thiếu:

```

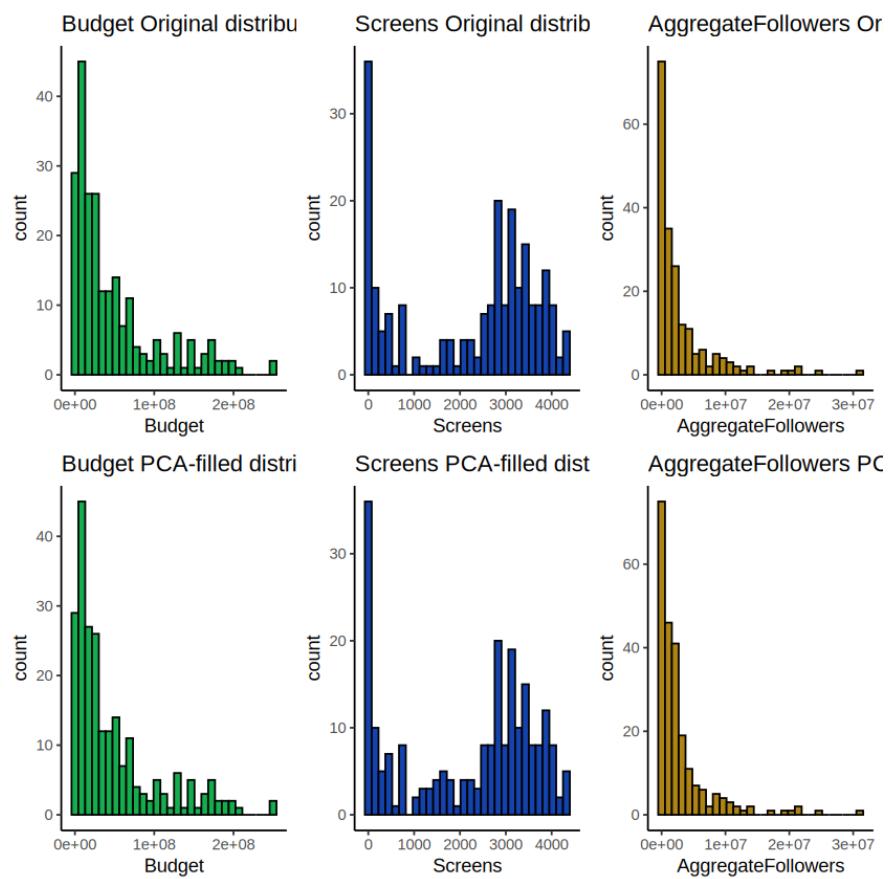
1 # Trực quan phân phối trước và sau khi fill missing value
2
3 h1 <- ggplot(raw_data, aes(x = Budget)) +
4   geom_histogram(fill = "#15ad4f", color = "#000000", position =
5     "identity") +
6   ggtitle("Budget Original distribution") +
7   theme_classic()
8
9 h2 <- ggplot(raw_data, aes(x = Screens)) +
10  geom_histogram(fill = "#1543ad", color = "#000000", position =
11    "identity") +
12    ggtitle("Screens Original distribution") +
13    theme_classic()
14
15
16
17 h3 <- ggplot(raw_data, aes(x = AggregateFollowers )) +
18   geom_histogram(fill = "#ad8415", color = "#000000", position =
19     "identity") +
20     ggtitle("AggregateFollowers Original distribution") +
21     theme_classic()
22
23
24
25 h4 <- ggplot(processed_data, aes(x = Budget)) +
26   geom_histogram(fill = "#15ad4f", color = "#000000", position =
27     "identity") +
28     ggtitle("Budget PCA-filled distribution") +
29     theme_classic()

```

```

21 h5 <- ggplot(processed_data, aes(x = Screens)) +
22   geom_histogram(fill = "#1543ad", color = "#000000", position
23     = "identity") +
24   ggtitle("Screens PCA-filled distribution") +
25   theme_classic()
26 h6 <- ggplot(processed_data, aes(x = AggregateFollowers )) +
27   geom_histogram(fill = "#ad8415", color = "#000000", position
28     = "identity") +
29   ggtitle("AggregateFollowers PCA-filled distribution") +
30   theme_classic()
31
32 plot_grid(h1, h2, h3, h4, h5, h6, nrow = 2, ncol = 3, rel_
33   widths = c(1, 1), rel_heights = c(1, 1))

```



Hình 2.40: Phân phối trước và sau khi điền các giá trị bị thiếu đối với các cột Budget, Screens và AggregateFollowers.

Nhận xét:

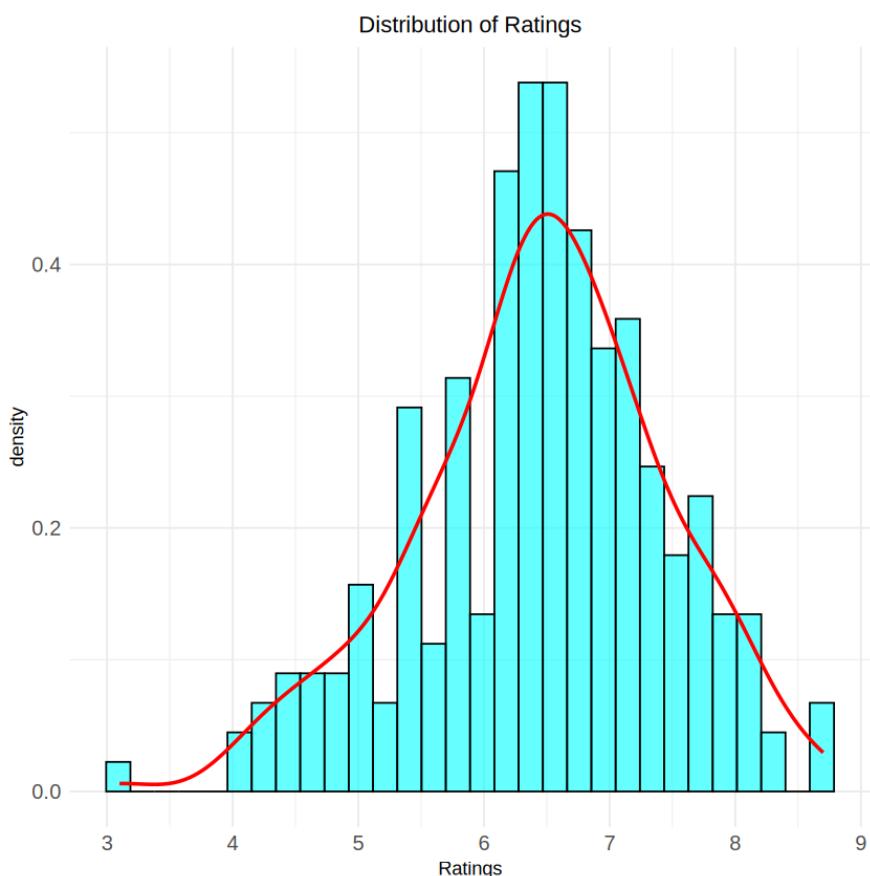
- Ta thấy phân phối sau khi điền không có quá nhiều chênh lệch so với phân phối dữ liệu gốc (đã bỏ qua các giá trị bị thiếu)

2.2.5. Quay lại bước khám phá và tiền xử lý dữ liệu

Trong phần này, chúng tôi khảo sát các biến để chuẩn hóa dữ liệu. Chúng tôi lựa chọn box-cox transformation để tìm kiếm giá trị λ tối ưu để biến đổi dữ liệu.

Phân tích biến Ratings

Trong phần này, chúng ta sẽ xem xét biến Ratings thể hiện điểm đánh giá đối với một bộ phim

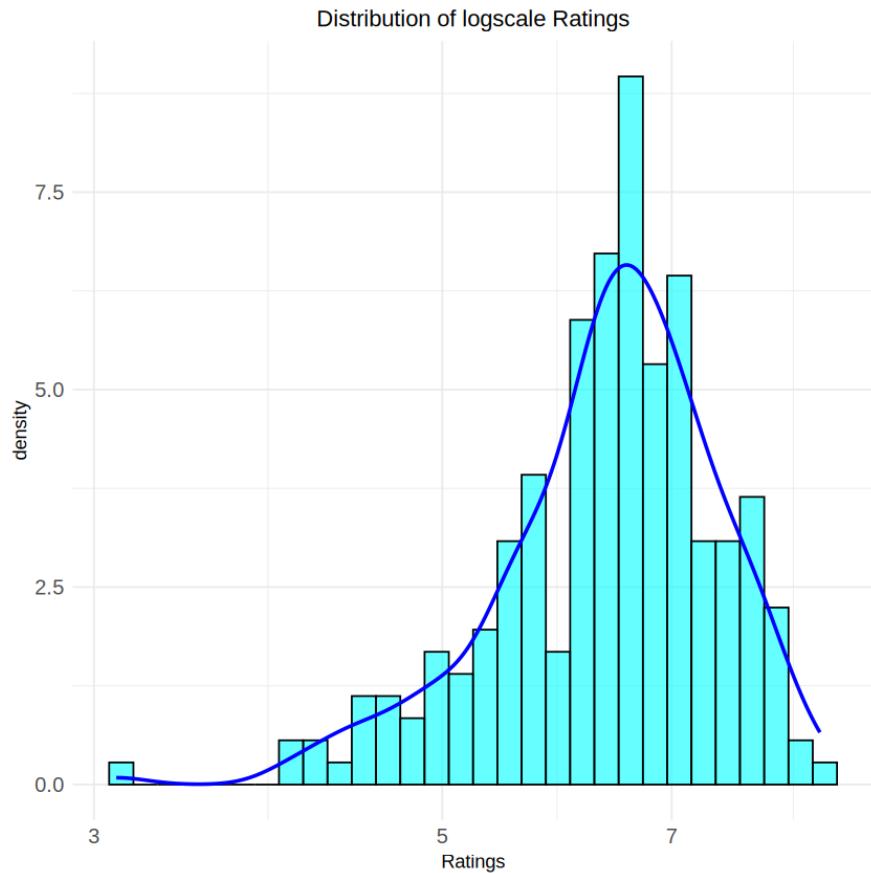


Hình 2.41: Phân phối ban đầu của Ratings.

Nhận xét:

- Nhìn vào biểu đồ, ta thấy phân phối của biến Ratings tương đối xấp xỉ chuẩn.

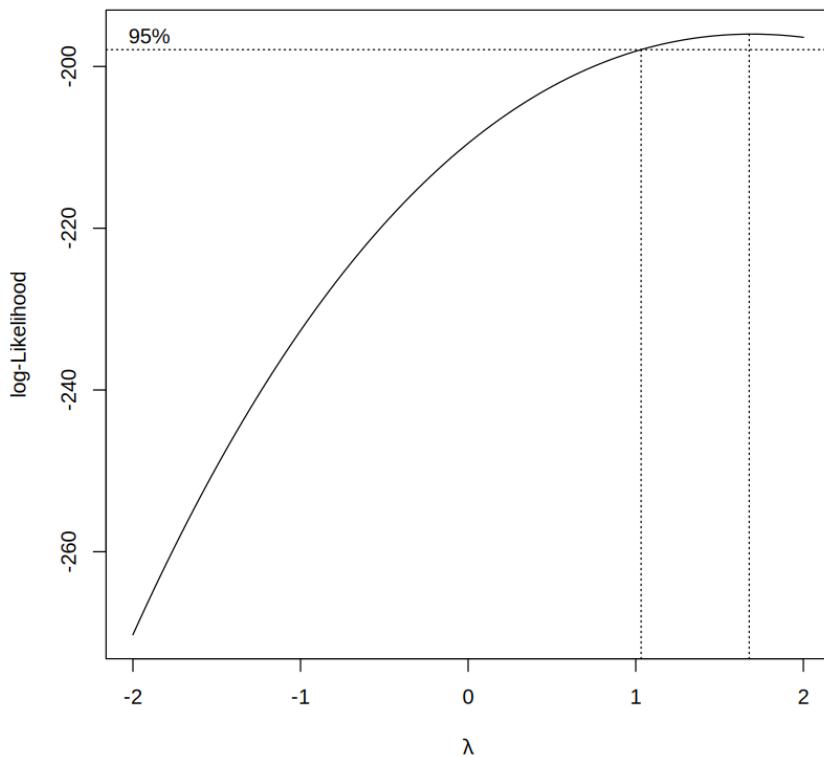
Ta thử sử dụng log-transform nó và thu được phân phối như hình bên dưới.



Hình 2.42: Phân phối sau khi log-scale của Ratings.

Nhận xét:

- Ta nhận thấy sau khi sử dụng log-transform, dữ liệu bị lệch trái (lệch âm). Do đó, ta thử sử dụng biến đổi box-cox.



Hình 2.43: Log-likelihood với các giá trị λ của Ratings.

Nhận xét:

- Dựa trên biểu đồ, ta tìm được giá trị lambda tối ưu với mức ý nghĩa 5% là 1.677.

Và ta thực hiện biến đổi dữ liệu với giá trị lambda vừa tìm được.



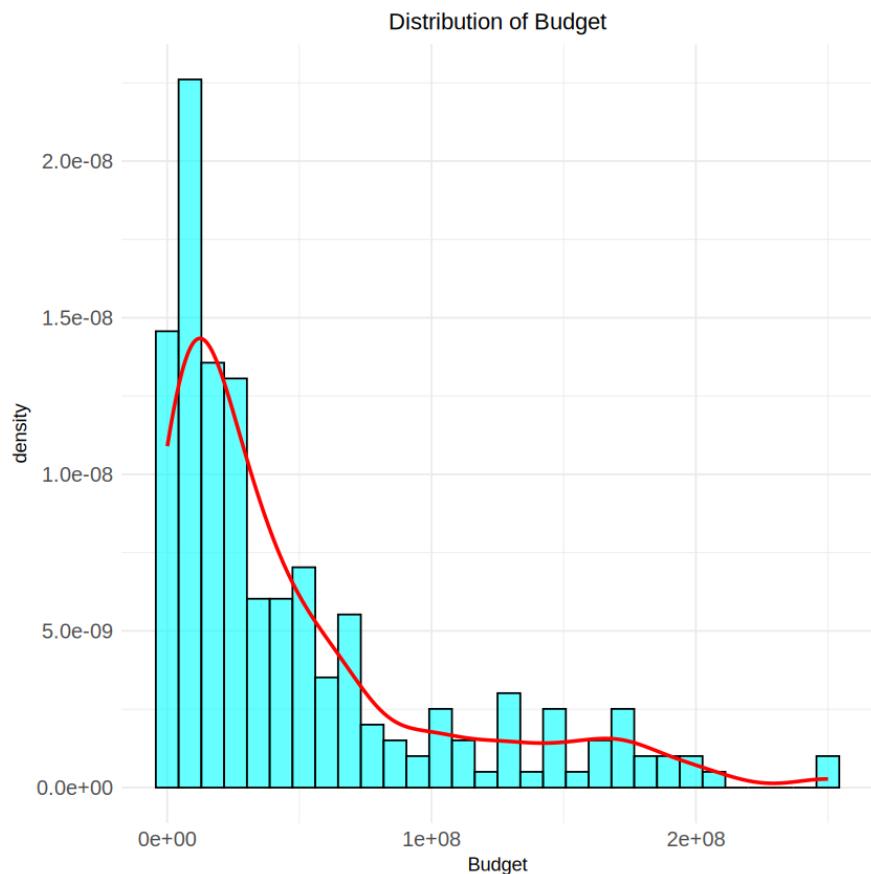
Hình 2.44: Phân phối trước và sau khi biến đổi của Gross.

Nhận xét:

- Ta có được giá trị lambda tối ưu là 1.677 và sử dụng giá trị này để biến đổi biến Ratings. Biểu đồ histogram phía bên dưới thể hiện phân phối của biến này trước và sau khi biến đổi. Dễ dàng thấy được, sau khi biến đổi, biến này đã tương đối chuẩn hơn.

Phân tích biến Budget

Trong phần này, chúng ta sẽ xem xét biến Budget thể hiện chi phí đầu tư cho một bộ phim.

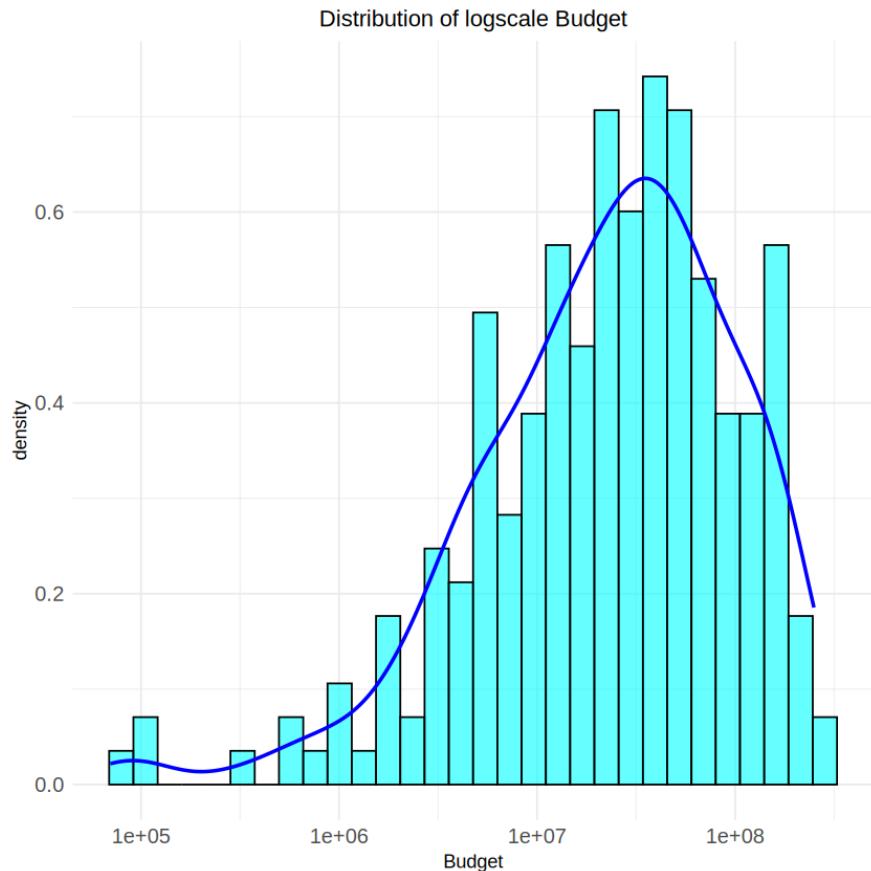


Hình 2.45: Phân phối ban đầu của Budget.

Nhận xét:

- Nhìn vào biểu đồ, ta thấy phân phối của biến Ratings bị lệch phải (lệch dương).

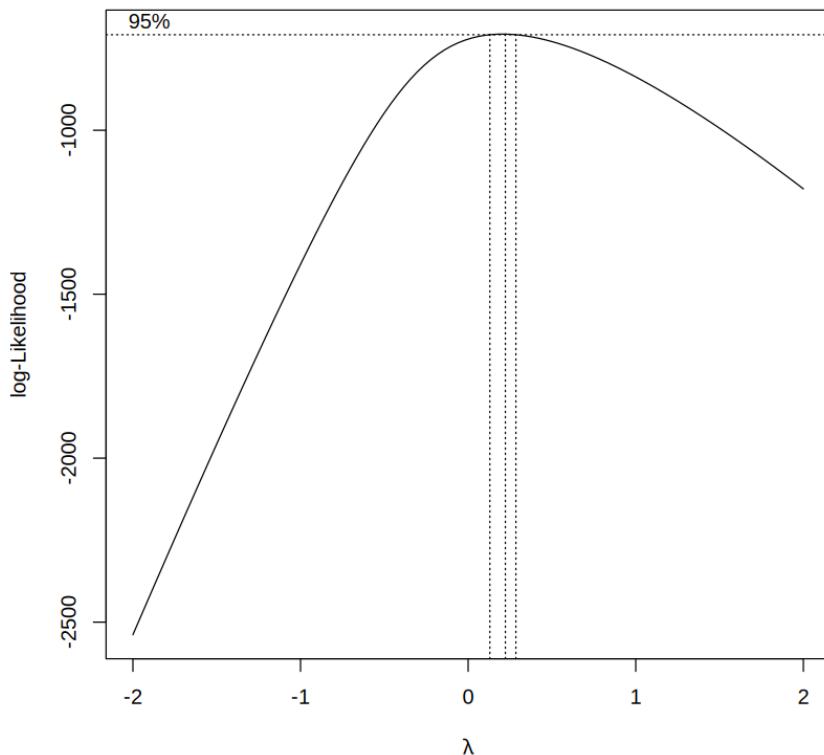
Ta thử sử dụng log-transform nó và thu được phân phối như hình bên dưới.



Hình 2.46: Phân phối sau khi log-scale của Budget.

Nhận xét:

- Ta nhận thấy sau khi sử dụng log-transform, dữ liệu tương đối xấp xỉ chuẩn. Do đó, ta thử sử dụng biến đổi box-cox.

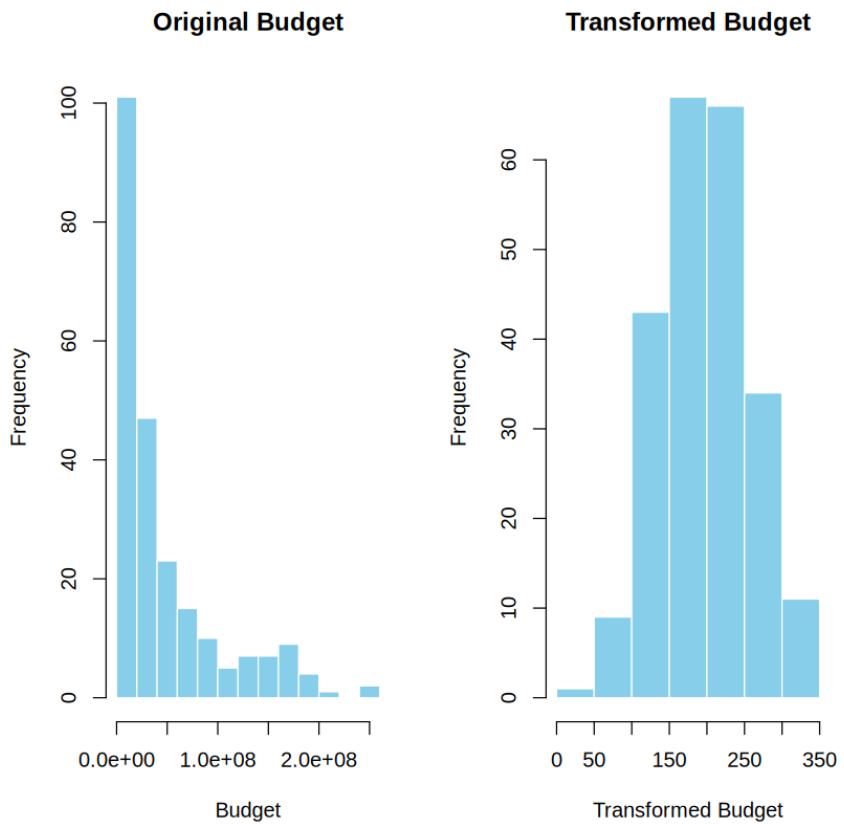


Hình 2.47: Log-likelihood với các giá trị λ của Budget.

Nhận xét:

- Dựa trên biểu đồ, ta tìm được giá trị lambda tối ưu với mức ý nghĩa 5% là 0.222.

Và ta thực hiện biến đổi dữ liệu với giá trị lambda vừa tìm được.



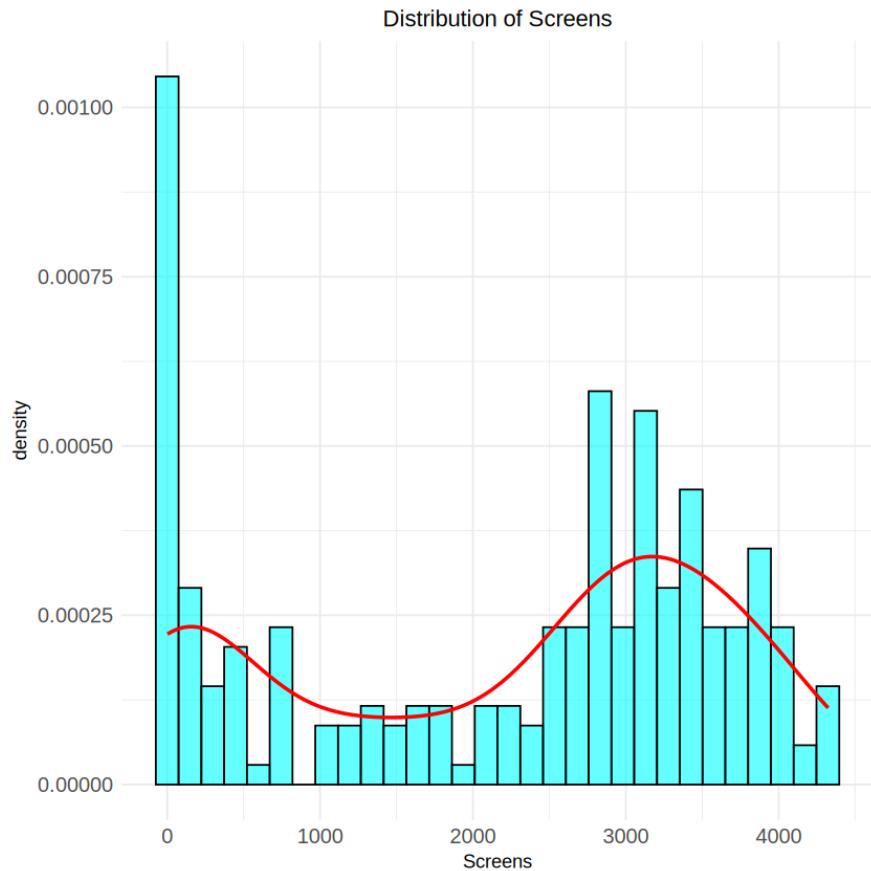
Hình 2.48: Phân phối trước và sau khi biến đổi của Budget.

Nhận xét:

- Ta có được giá trị lambda tối ưu là 0.2222 và sử dụng giá trị này để biến đổi biến Budget. Biểu đồ histogram phía bên dưới thể hiện phân phối của biến này trước và sau khi biến đổi. Dễ dàng thấy được, sau khi biến đổi, biến này đã tương đối chuẩn hơn.

Phân tích biến Screens

Trong phần này, chúng ta sẽ xem xét biến Screens thể hiện số rạp chiếu của một bộ phim.

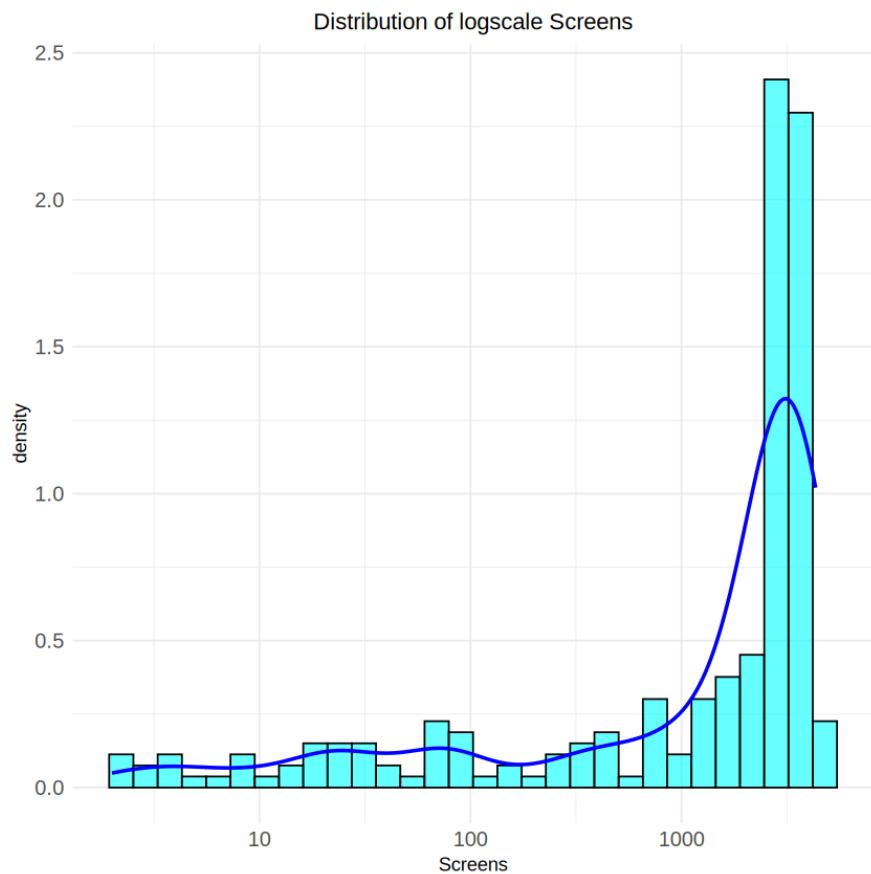


Hình 2.49: Phân phối ban đầu của Screens.

Nhận xét:

- Nhìn vào biểu đồ, ta thấy phân phối của biến Screens có phân phối hai đỉnh, trong đó một đỉnh tập trung ở gần 0 và một đỉnh tập trung ở 3000.

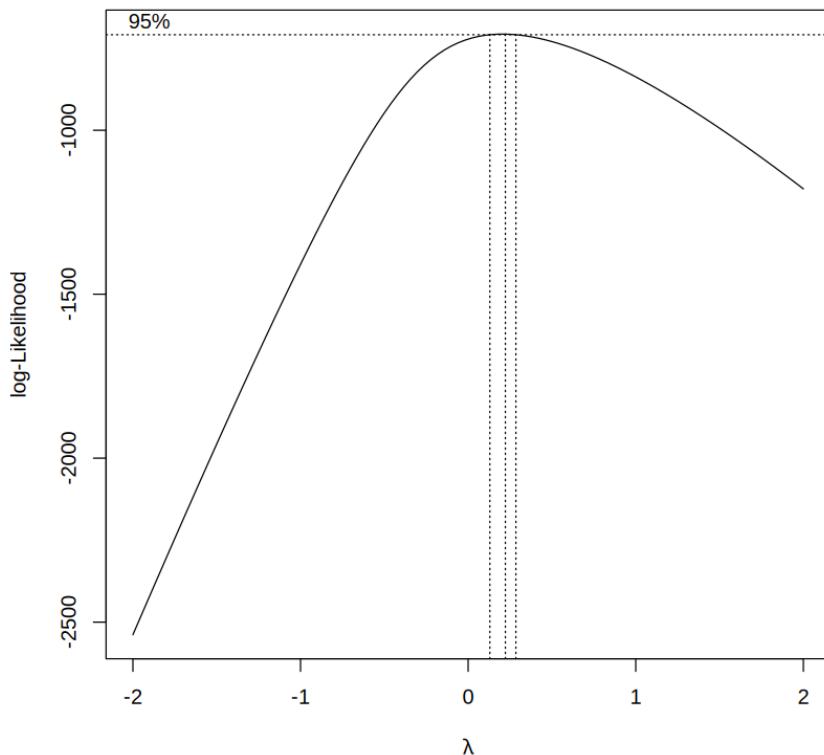
Ta thử sử dụng log-transform nó và thu được phân phối như hình bên dưới.



Hình 2.50: Phân phối sau khi log-scale của Screens.

Nhận xét:

- Ta nhận thấy sau khi sử dụng log-transform, dữ liệu tương đối bị lệch phải (lệch dương). Do đó, ta thử sử dụng biến đổi box-cox.

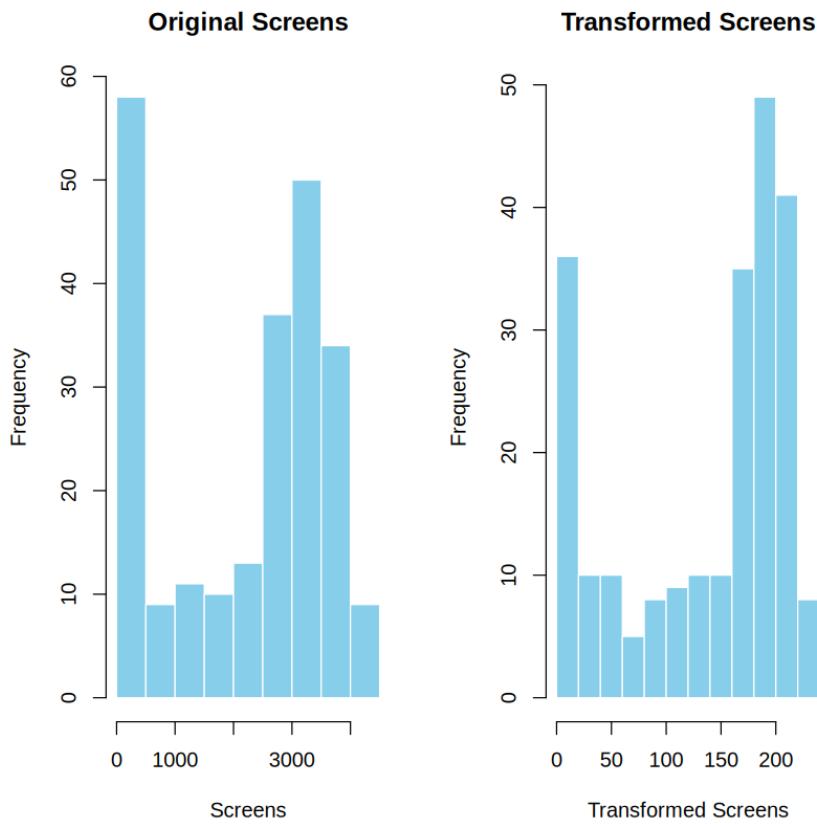


Hình 2.51: Log-likelihood với các giá trị λ của Screens.

Nhận xét:

- Dựa trên biểu đồ, ta tìm được giá trị lambda tối ưu với mức ý nghĩa 5% là 0.5858.

Và ta thực hiện biến đổi dữ liệu với giá trị lambda vừa tìm được.



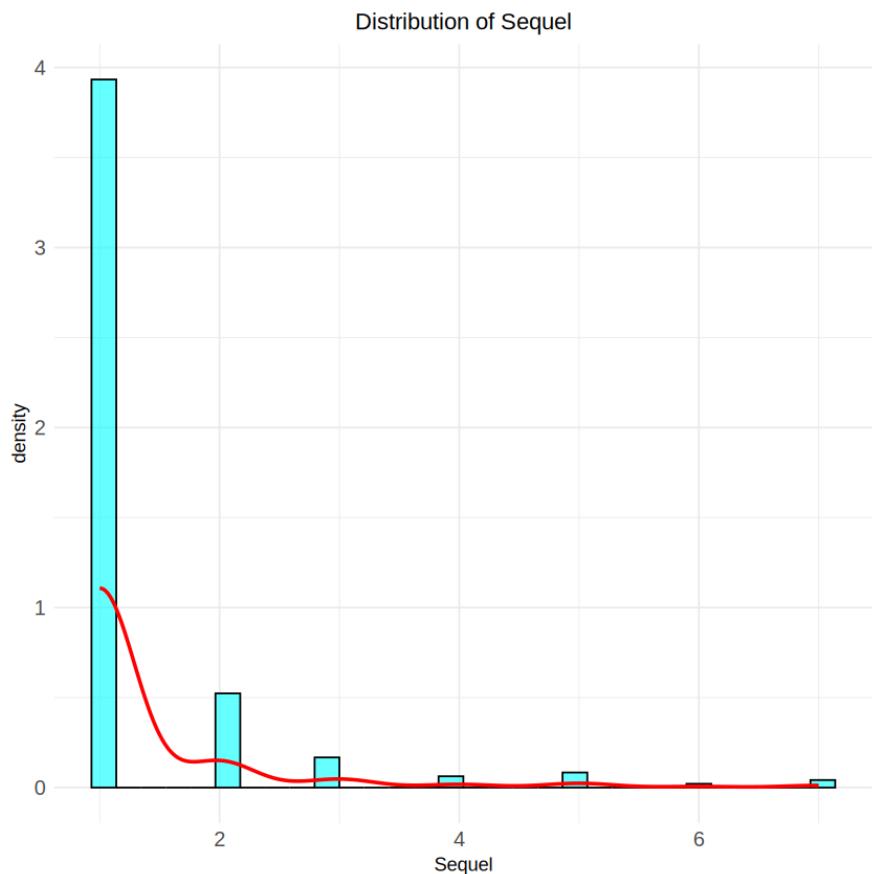
Hình 2.52: Phân phối trước và sau khi biến đổi của Screens.

Nhận xét:

- Ta có được giá trị lambda tối ưu là 0.2222 và sử dụng giá trị này để biến đổi biến Screens. Biểu đồ histogram phía bên dưới thể hiện phân phối của biến này trước và sau khi biến đổi. Dễ dàng thấy được, sau khi biến đổi, biến này vẫn có phân phối hai đỉnh trong đó một đỉnh tập trung ở gần 0 còn đỉnh còn lại tập trung ở 200.
- Như vậy, có thể có ngoại lai xuất hiện, ta cần thực hiện loại bỏ các cực ngoại lai để tiếp tục xây dựng mô hình.

Phân tích biến Sequel

Trong phần này, chúng ta sẽ xem xét biến Sequel thể hiện các phần của một bộ phim.



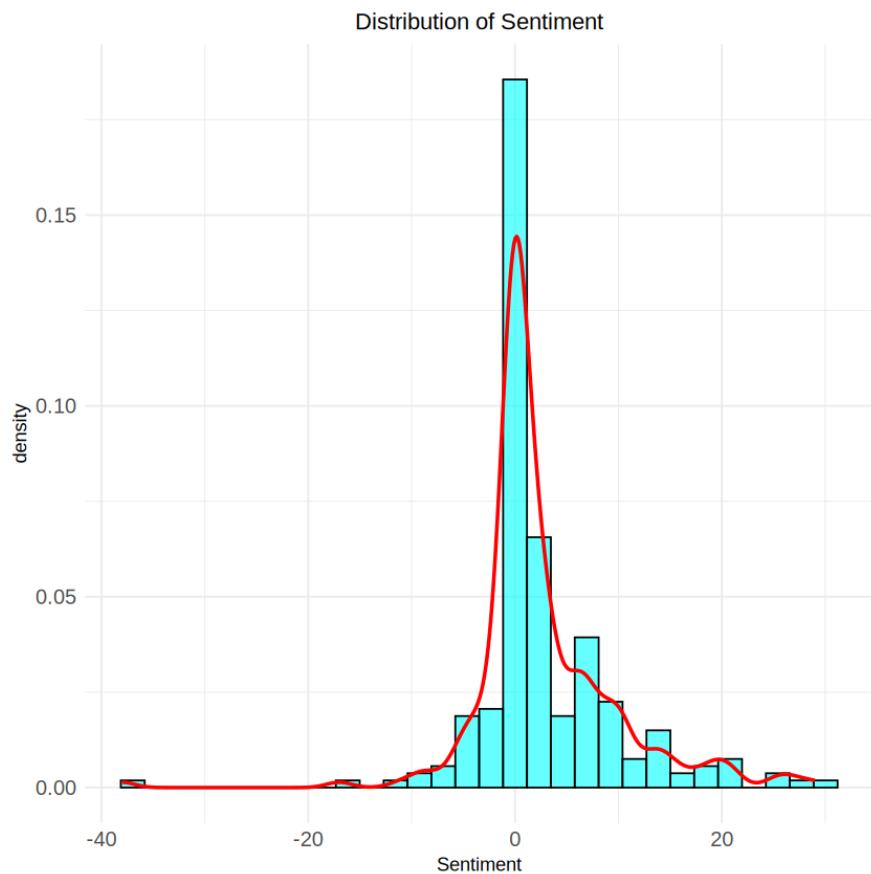
Hình 2.53: Phân phối ban đầu của Sequel.

Nhận xét:

- Nhìn vào biểu đồ, ta thấy phân phối của biến Ratings bị lệch trái (lệch âm).
- Đa số các bộ phim chỉ có một phần phim
- Rất ít các bộ phim có từ 4 phần trở lên
- Số lượng các bộ phim có 5 phần nhiều hơn các bộ phim có 4 phần, và 6 phần

Phân tích biến Sentiment

Trong phần này, chúng ta sẽ xem xét biến Sentiment thể hiện ý kiến khán giả đối với một bộ phim.

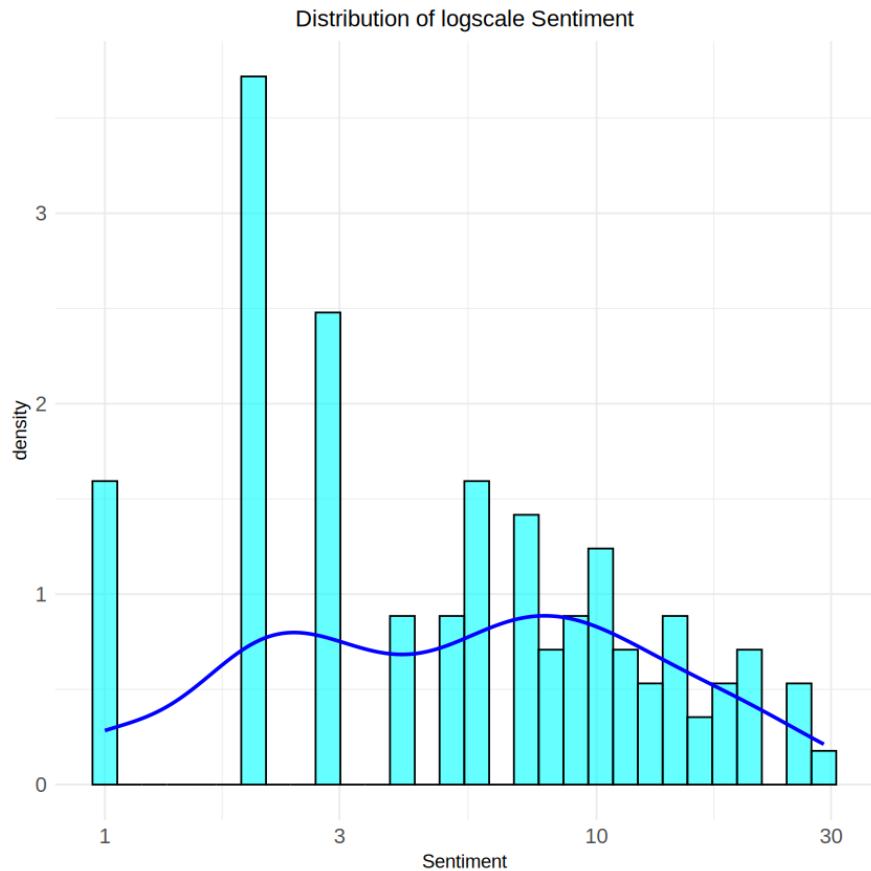


Hình 2.54: Phân phối ban đầu của Sentiment.

Nhận xét:

- Nhìn vào biểu đồ, ta thấy phân phối của biến Sentiment có phân phối xấp xỉ chuẩn.
- Tuy nhiên, các giá trị tập trung ở 0 rất cao.

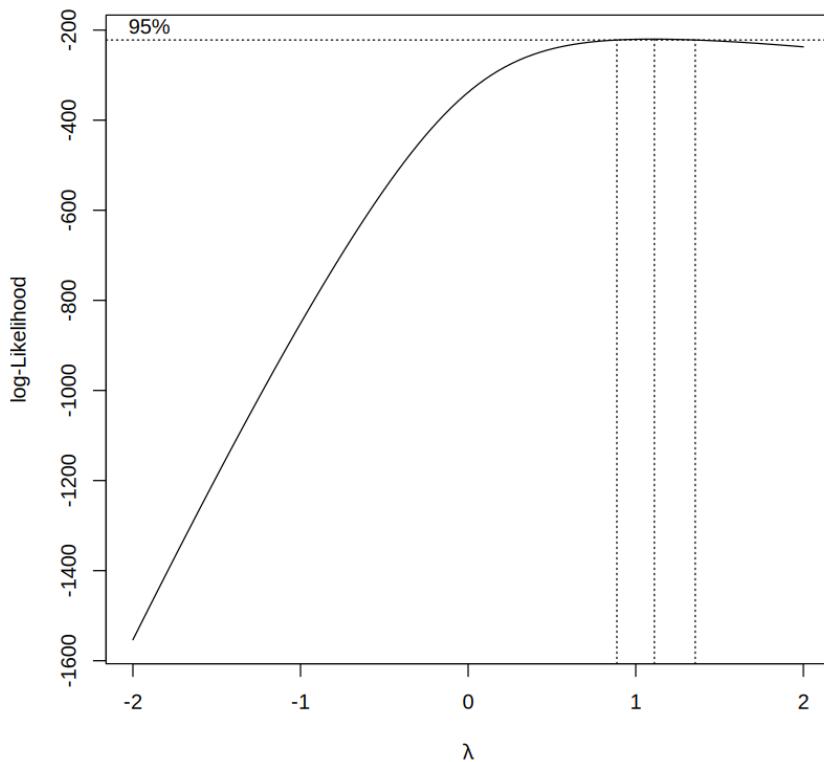
Ta thử sử dụng log-transform nó và thu được phân phối như hình bên dưới.



Hình 2.55: Phân phối sau khi log-scale của Sentiment.

Nhận xét:

- Ta nhận thấy sau khi sử dụng log-transform, dữ liệu tương đối lệch trái (lệch dương). Do đó, ta thử sử dụng biến đổi box-cox.

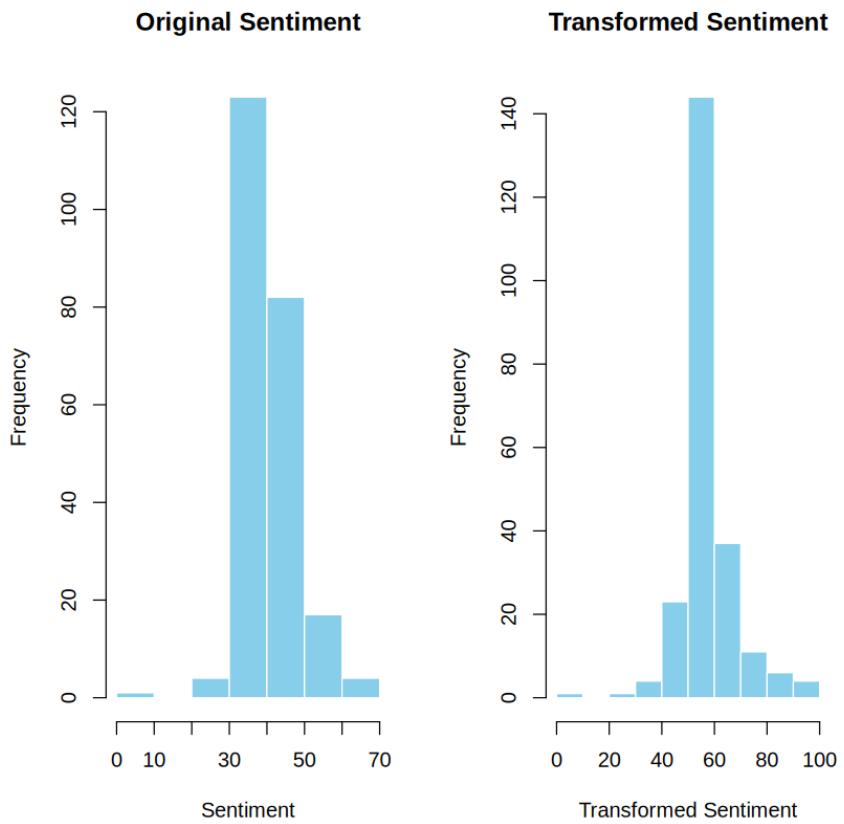


Hình 2.56: Log-likelihood với các giá trị λ của Sentiment.

Nhận xét:

- Dựa trên biểu đồ, ta tìm được giá trị lambda tối ưu với mức ý nghĩa 5% là 1.111.

Và ta thực hiện biến đổi dữ liệu với giá trị lambda vừa tìm được.



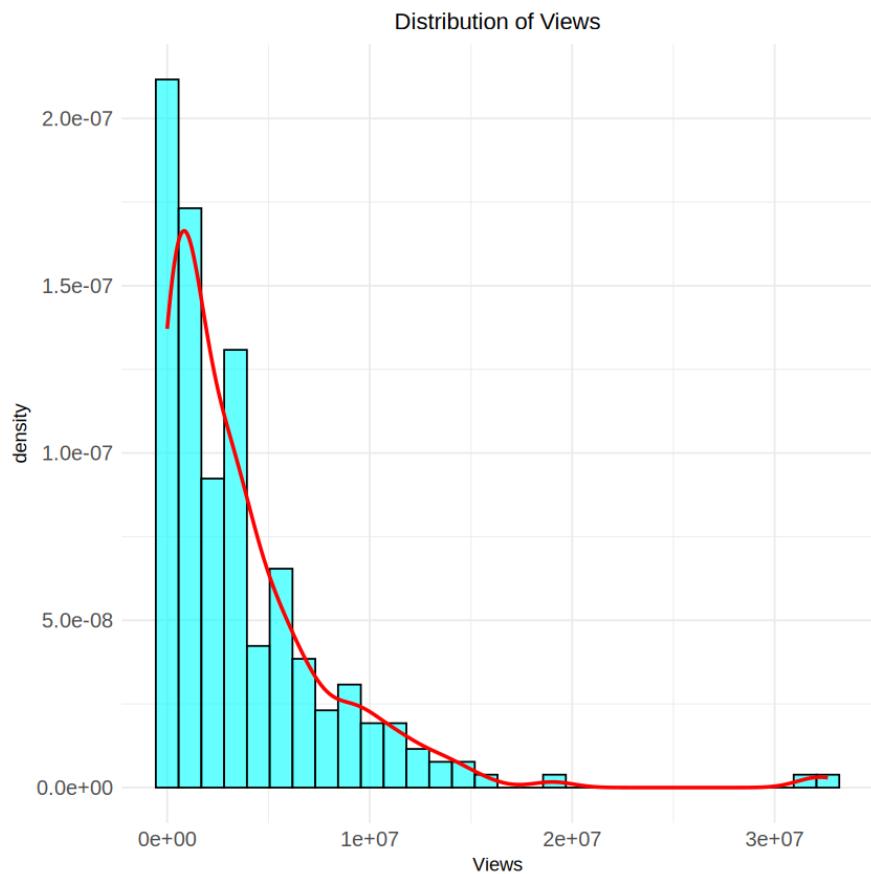
Hình 2.57: Phân phối trước và sau khi biến đổi của Sentiment.

Nhận xét:

- Ta có được giá trị lambda tối ưu là 1.11 và sử dụng giá trị này để biến đổi biến Sentiment. Biểu đồ histogram phía bên dưới thể hiện phân phối của biến này trước và sau khi biến đổi. Dễ dàng thấy được, sau khi biến đổi, biến này đã tương đối chuẩn hơn.

Phân tích biến Views

Trong phần này, chúng ta sẽ xem xét biến Views thể hiện số lượt xem của một bộ phim.

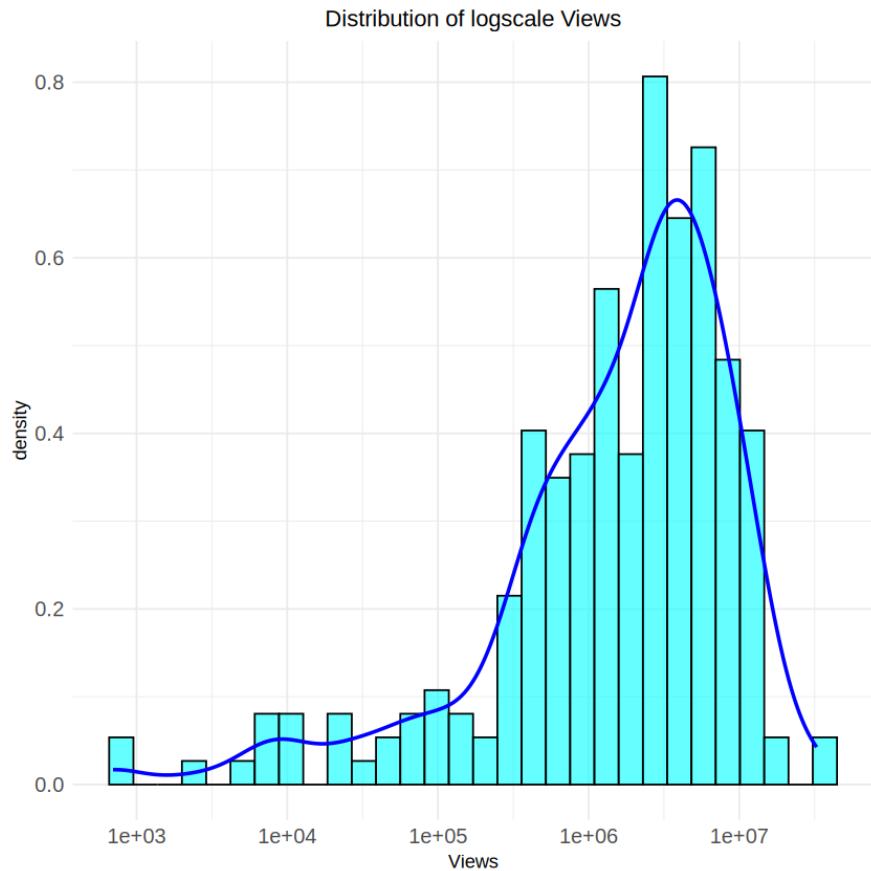


Hình 2.58: Phân phối ban đầu của Views.

Nhận xét:

- Nhìn vào biểu đồ, ta thấy phân phối của biến Views có phân phối bị lệch trái (lệch dương).

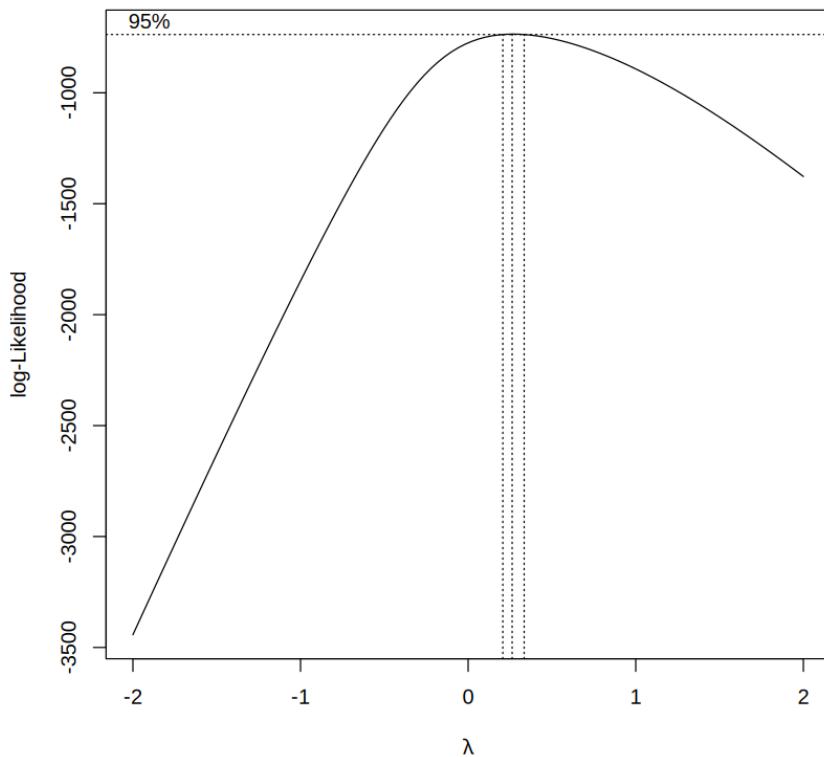
Ta thử sử dụng log-transform nó và thu được phân phối như hình bên dưới.



Hình 2.59: Phân phối sau khi log-scale của Views.

Nhận xét:

- Ta nhận thấy sau khi sử dụng log-transform, dữ liệu tương đối lệch phải (lệch âm). Do đó, ta thử sử dụng biến đổi box-cox.

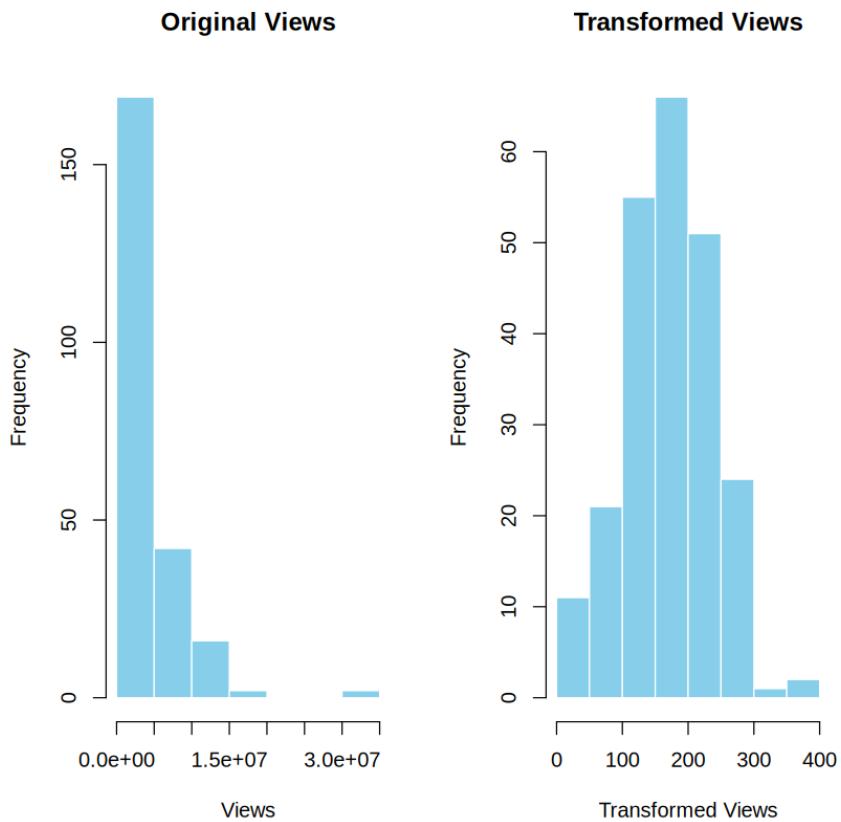


Hình 2.60: Log-likelihood với các giá trị λ của Views.

Nhận xét:

- Dựa trên biểu đồ, ta tìm được giá trị lambda tối ưu với mức ý nghĩa 5% là 0.2626.

Và ta thực hiện biến đổi dữ liệu với giá trị lambda vừa tìm được.



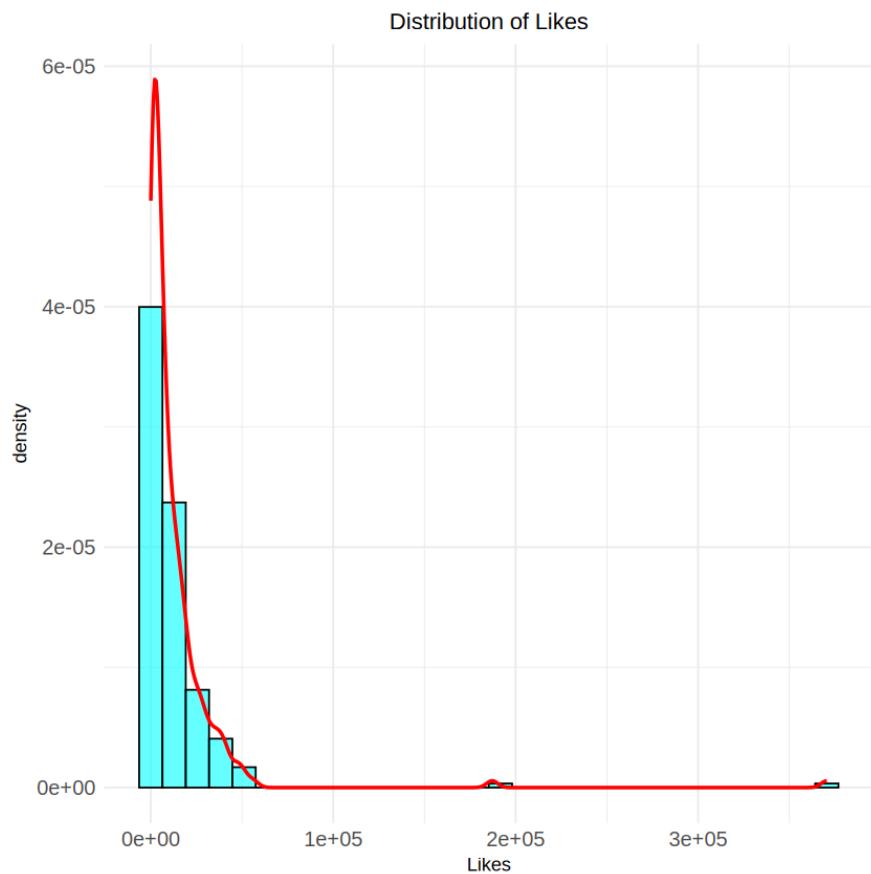
Hình 2.61: Phân phối trước và sau khi biến đổi của Views.

Nhận xét:

- Ta có được giá trị lambda tối ưu là 0.2626 và sử dụng giá trị này để biến đổi biến Views. Biểu đồ histogram phía bên dưới thể hiện phân phối của biến này trước và sau khi biến đổi. Dễ dàng thấy được, sau khi biến đổi, biến này đã tương đối chuẩn hơn.

Phân tích biến Likes

Trong phần này, chúng ta sẽ xem xét biến Likes thể hiện số lượt thích của một bộ phim.

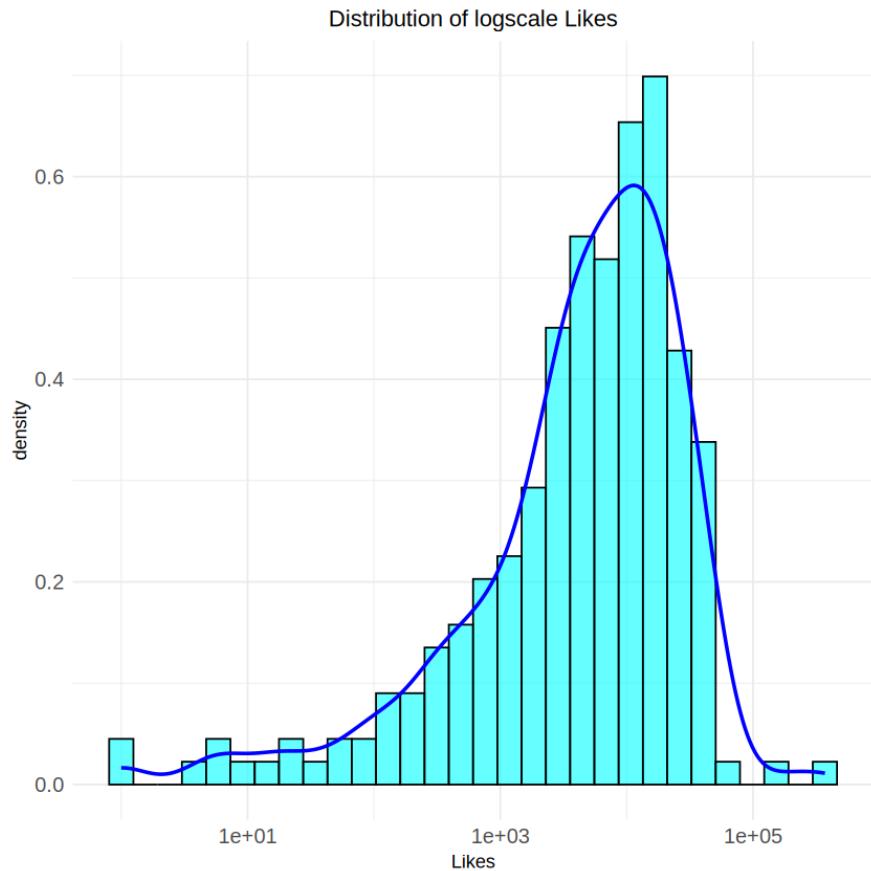


Hình 2.62: Phân phối ban đầu của Likes.

Nhận xét:

- Nhìn vào biểu đồ, ta thấy phân phối của biến Likes có phân phối bị lệch trái (lệch dương).

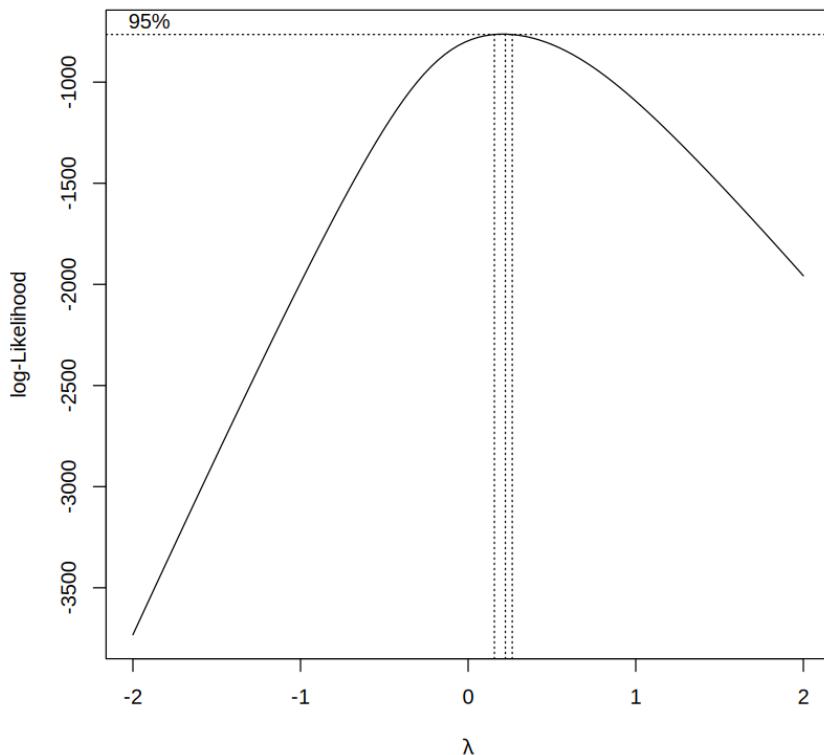
Ta thử sử dụng log-transform nó và thu được phân phối như hình bên dưới.



Hình 2.63: Phân phối sau khi log-scale của Likes.

Nhận xét:

- Ta nhận thấy sau khi sử dụng log-transform, dữ liệu tương đối lệch phải (lệch âm). Do đó, ta thử sử dụng biến đổi box-cox.

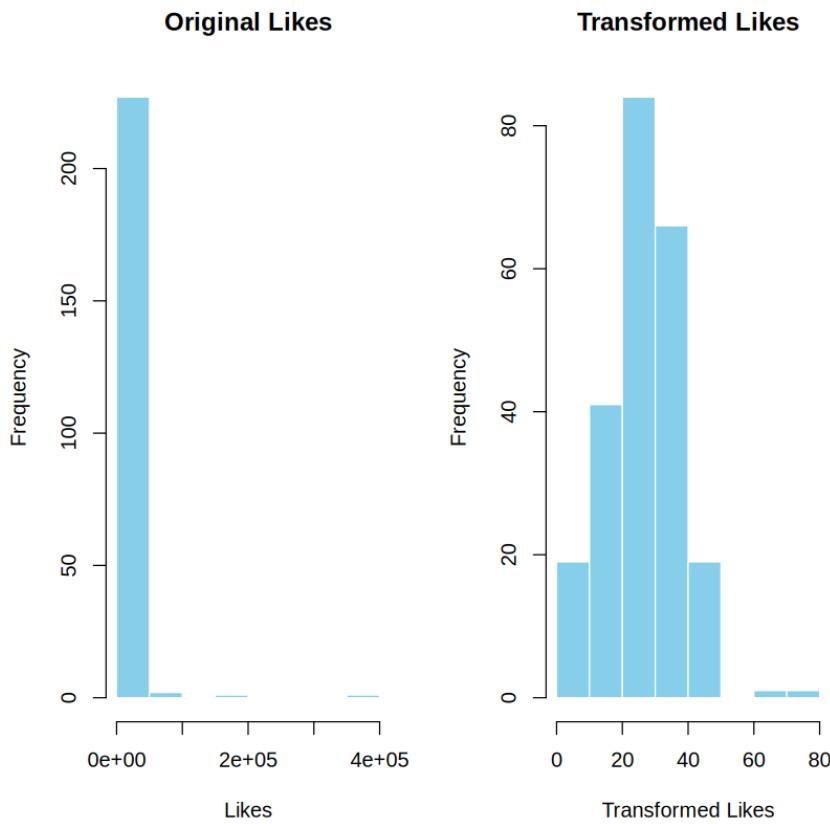


Hình 2.64: Log-likelihood với các giá trị λ của Likes.

Nhận xét:

- Dựa trên biểu đồ, ta tìm được giá trị lambda tối ưu với mức ý nghĩa 5% là 0.2222.

Và ta thực hiện biến đổi dữ liệu với giá trị lambda vừa tìm được.



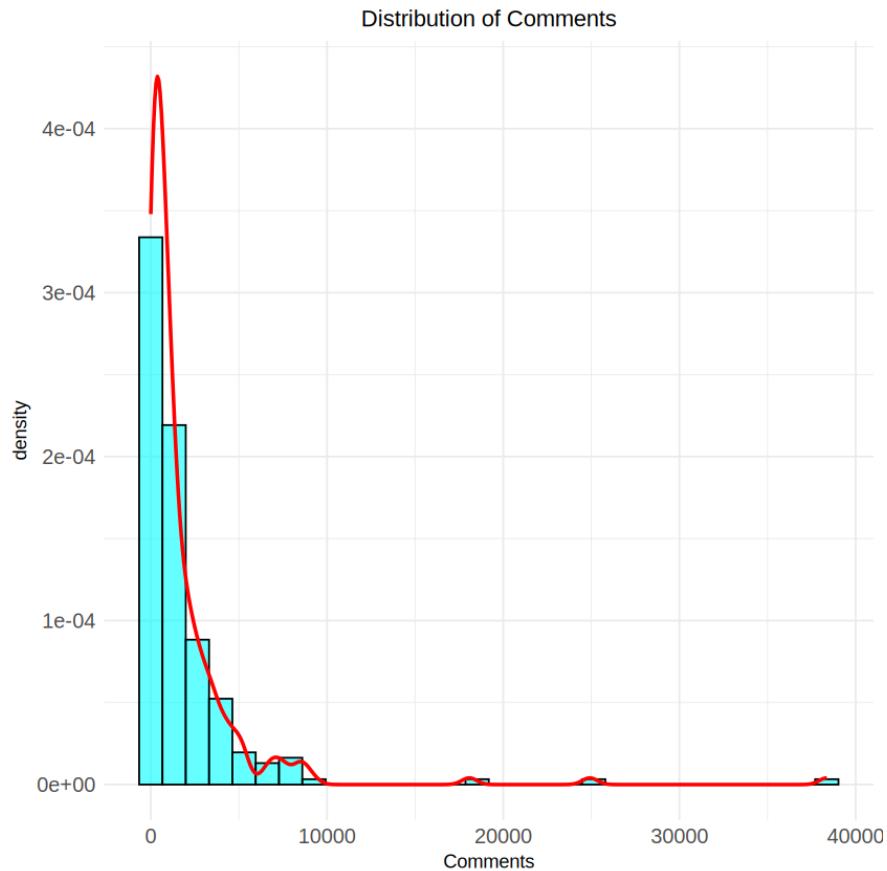
Hình 2.65: Phân phối trước và sau khi biến đổi của Likes.

Nhận xét:

- Ta có được giá trị lambda tối ưu là 0.2222 và sử dụng giá trị này để biến đổi biến Likes. Biểu đồ histogram phía bên dưới thể hiện phân phối của biến này trước và sau khi biến đổi. Dễ dàng thấy được, sau khi biến đổi, biến này đã tương đối chuẩn hơn.

Phân tích biến Comments

Trong phần này, chúng ta sẽ xem xét biến Comments thể hiện số bình luận của một bộ phim.

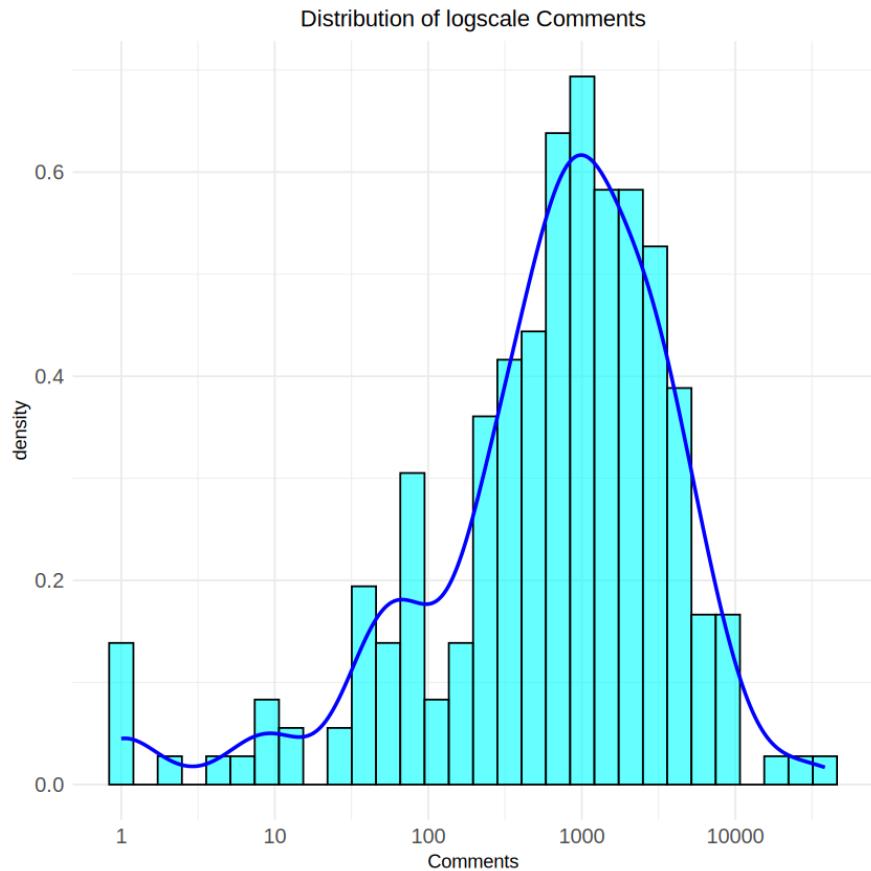


Hình 2.66: Phân phối ban đầu của Likes.

Nhận xét:

- Nhìn vào biểu đồ, ta thấy phân phối của biến Comments có phân phối bị lệch trái (lệch dương).

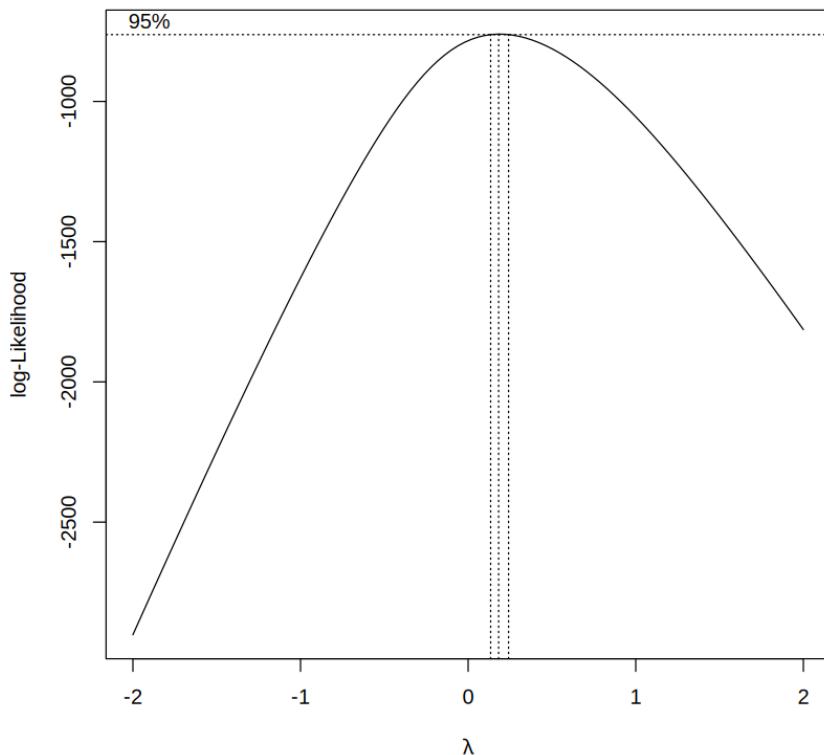
Ta thử sử dụng log-transform nó và thu được phân phối như hình bên dưới.



Hình 2.67: Phân phối sau khi log-scale của Comments.

Nhận xét:

- Ta nhận thấy sau khi sử dụng log-transform, dữ liệu tương đối lệch phải (lệch âm). Do đó, ta thử sử dụng biến đổi box-cox.

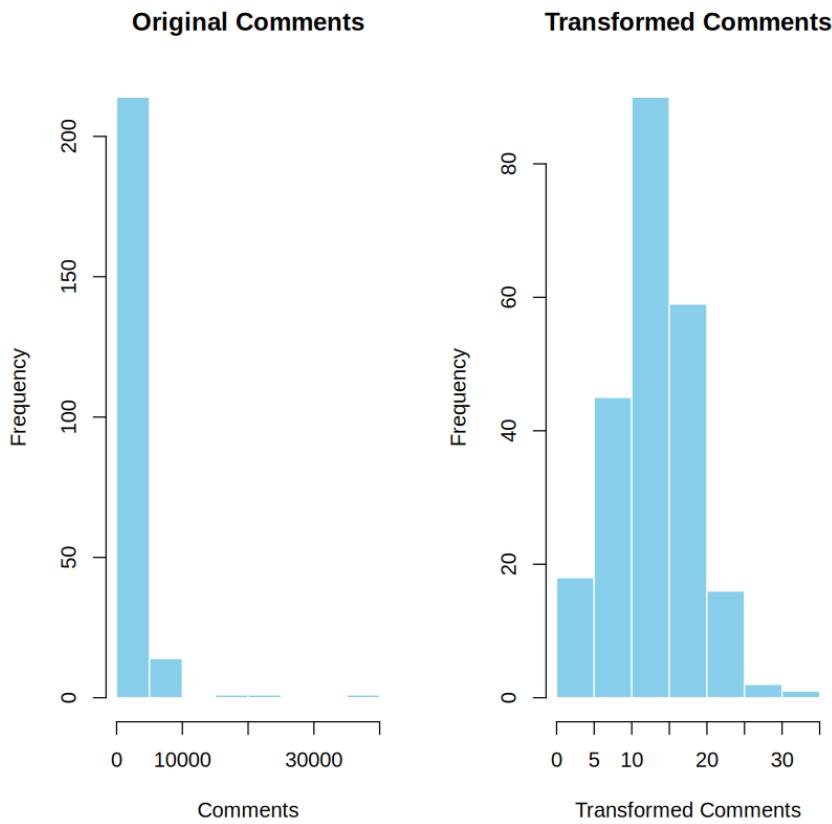


Hình 2.68: Log-likelihood với các giá trị λ của Comments.

Nhận xét:

- Dựa trên biểu đồ, ta tìm được giá trị lambda tối ưu với mức ý nghĩa 5% là 0.1818.

Và ta thực hiện biến đổi dữ liệu với giá trị lambda vừa tìm được.

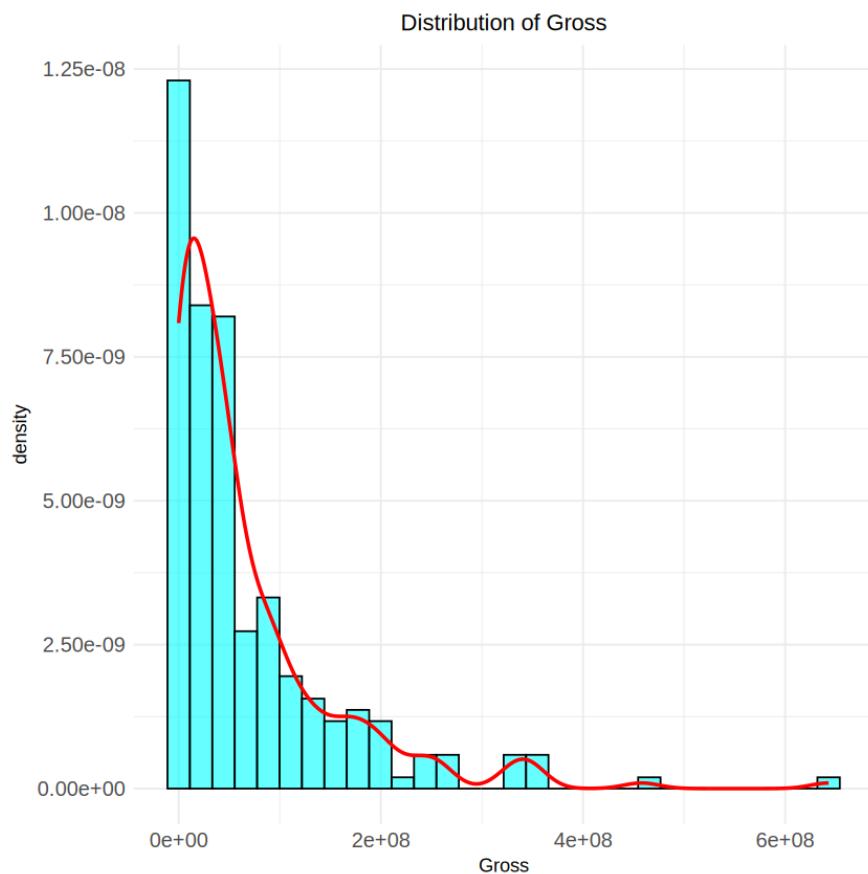


Hình 2.69: Phân phối trước và sau khi biến đổi của Comments.

Nhận xét:

- Ta có được giá trị lambda tối ưu là 0.1818 và sử dụng giá trị này để biến đổi biến Comments. Biểu đồ histogram phía bên dưới thể hiện phân phối của biến này trước và sau khi biến đổi. Dễ dàng thấy được, sau khi biến đổi, biến này đã tương đối chuẩn hơn.

Phân tích biến Gross

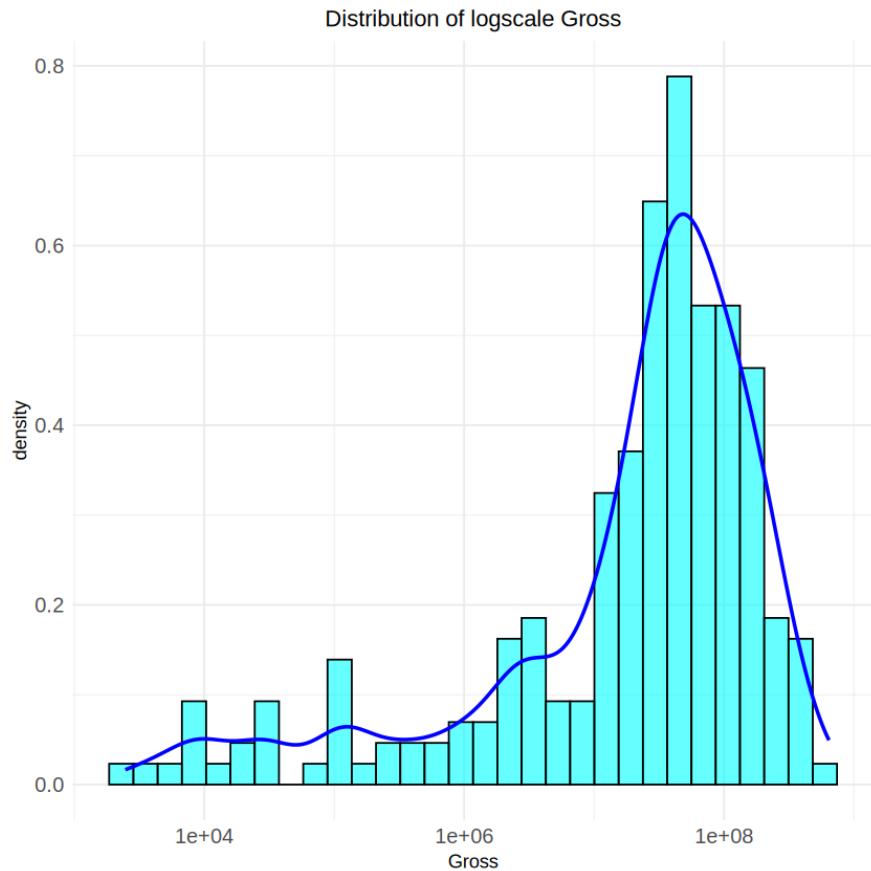


Hình 2.70: Phân phối ban đầu của Gross.

Nhận xét:

- Nhìn vào biểu đồ, ta thấy phân phối của biến Gross bị lệch trái.

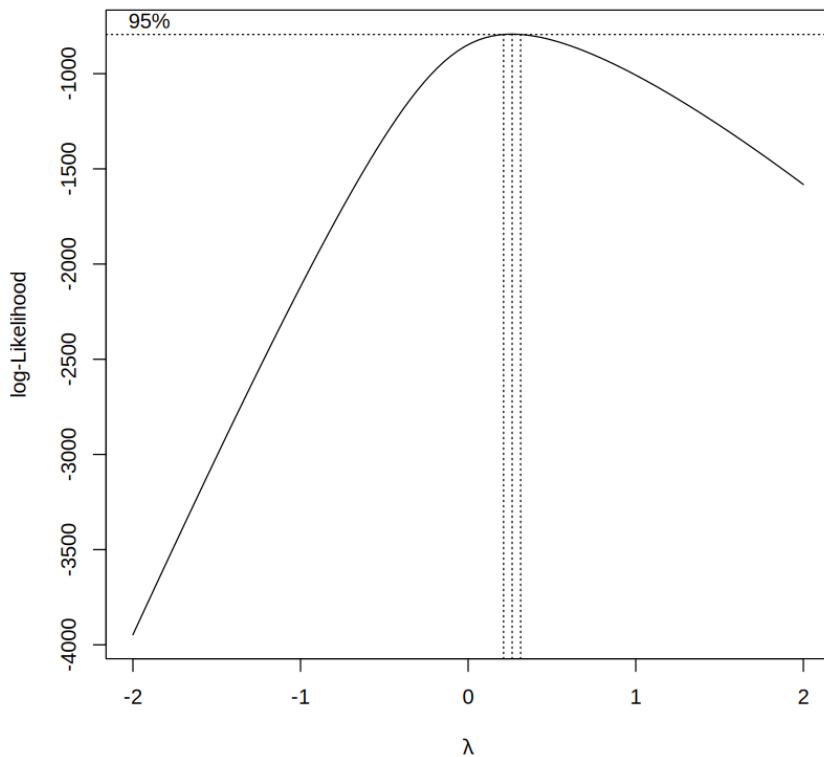
Ta thử sử dụng log-transform nó.



Hình 2.71: Phân phối sau khi log-scale của Gross.

Nhận xét:

- Ta nhận thấy sau khi sử dụng log-transform, dữ liệu bị lệch phải. Do đó, ta thử sử dụng box-cox.

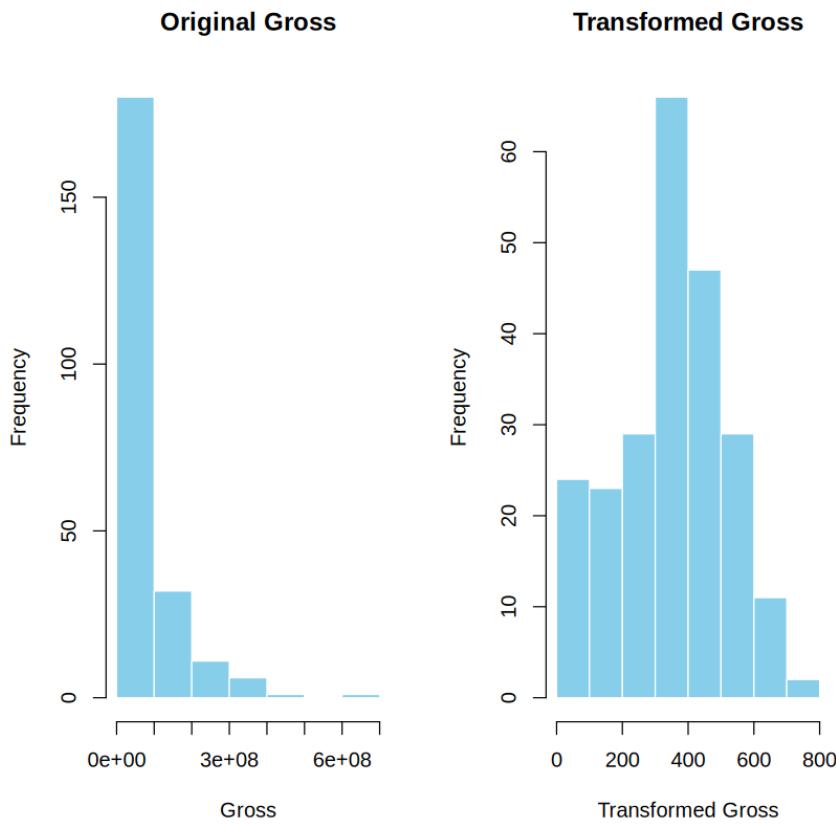


Hình 2.72: Log-likelihood với các giá trị λ của Gross.

Nhận xét:

- Dựa trên biểu đồ, ta tìm được giá trị lambda tối ưu với mức ý nghĩa 5% là 0.262

Và ta thực hiện biến đổi dữ liệu với giá trị lambda vừa tìm được.

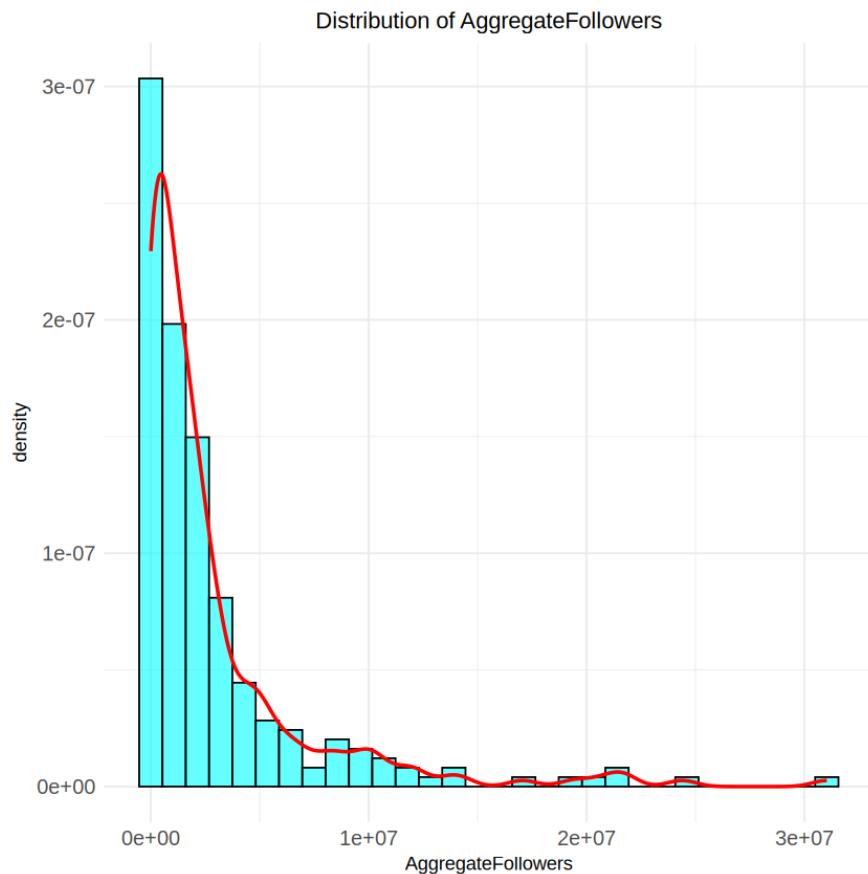


Hình 2.73: Phân phối trước và sau khi biến đổi của Gross.

Nhận xét:

- Ta có được giá trị lambda tối ưu là 0.262 và sử dụng giá trị này để biến đổi biến Gross. Biểu đồ histogram phía bên dưới thể hiện phân phối của biến này trước và sau khi biến đổi. Dễ dàng thấy được, sau khi biến đổi, biến này đã tương đối chuẩn hơn.

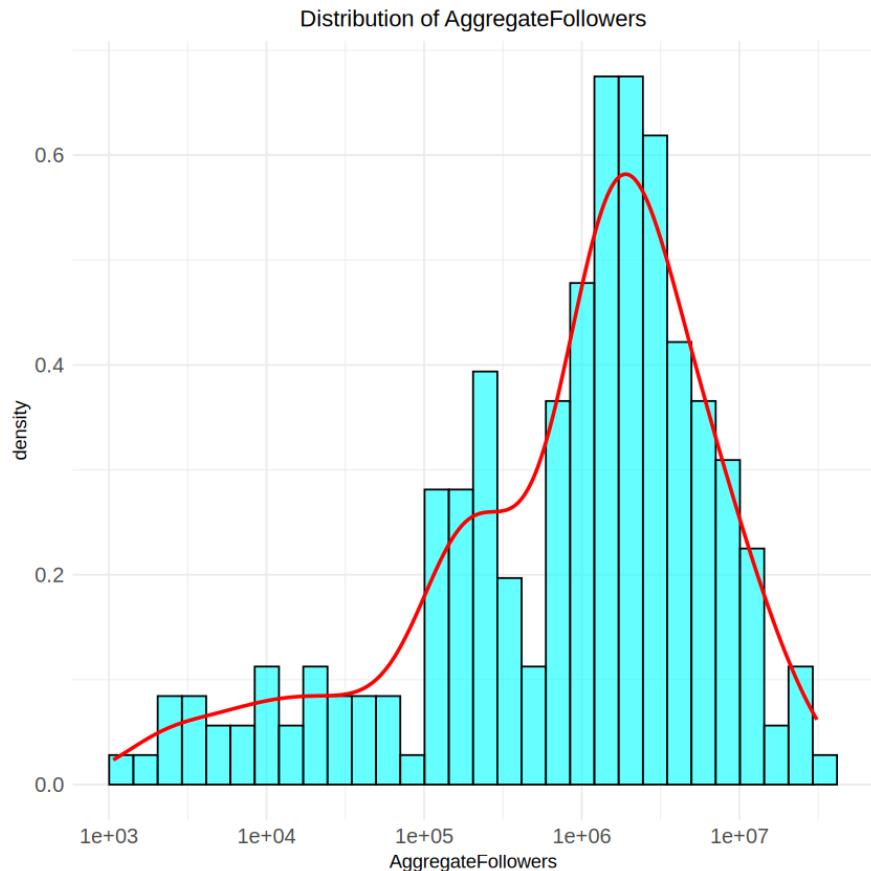
Phân tích biến AggregateFollowers



Hình 2.74: Phân phối ban đầu của Gross.

Nhận xét:

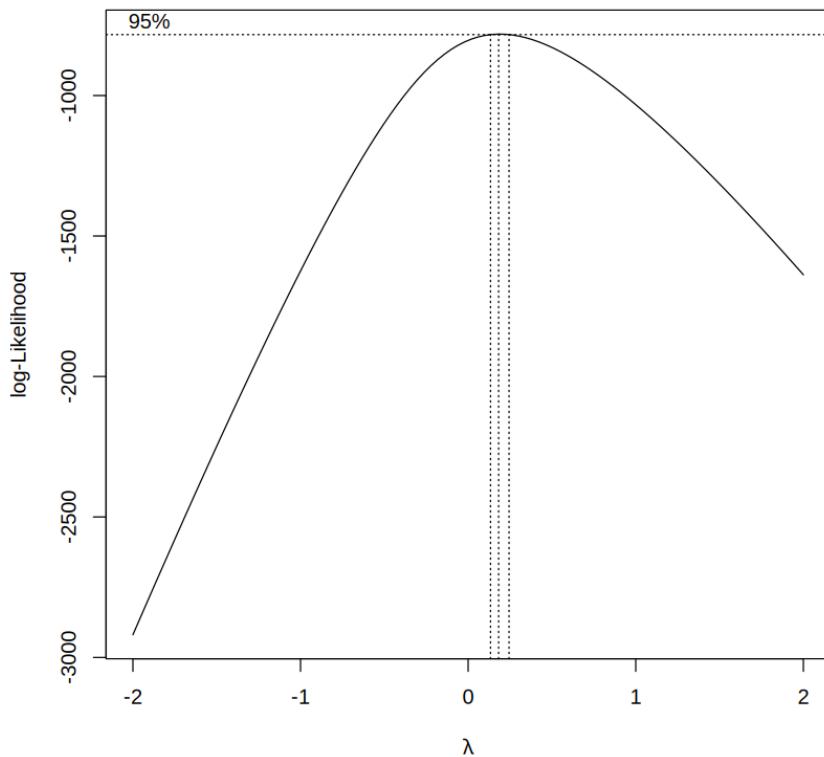
- Nhìn vào biểu đồ, ta thấy phân phối của biến AggregateFollowers bị lệch phải.



Hình 2.75: Phân phối sau khi logscale của AggregateFollowers.

Nhận xét:

- Khi dùng log-scale, phân phối của biến Comments đã xấp xỉ chuẩn hơn.
- Ta có thể dùng box-cox để biến đổi dữ liệu nhờ vào việc tìm lambda tối ưu.

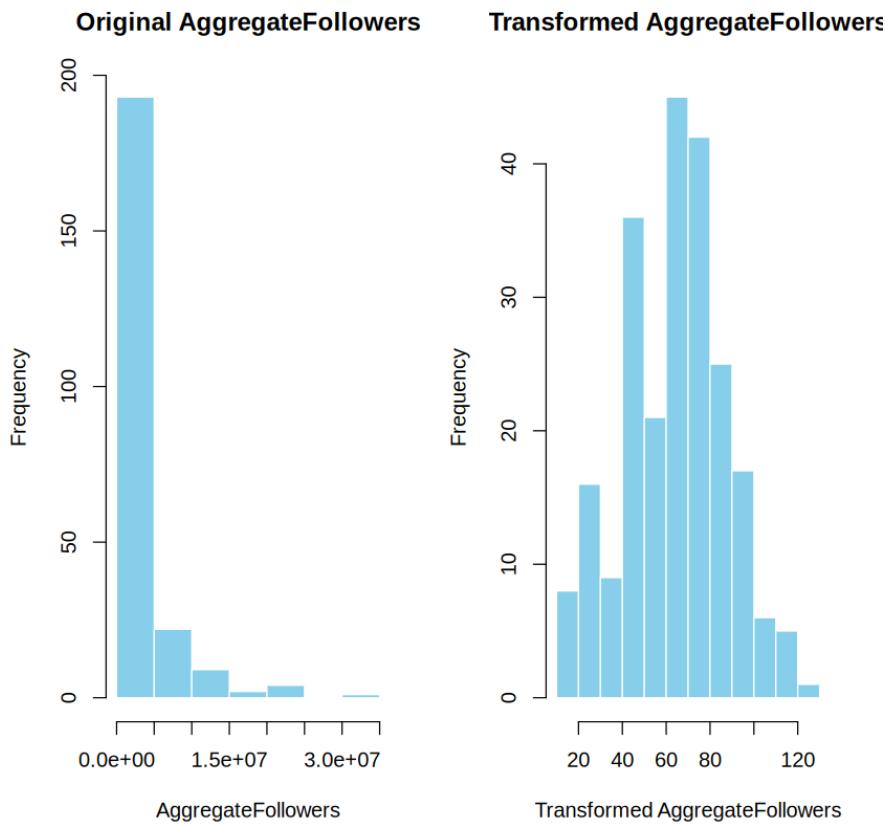


Hình 2.76: Log-likelihood với các giá trị λ của AggregateFollowers.

Nhận xét:

- Dựa trên biểu đồ, ta tìm được giá trị lambda tối ưu với mức ý nghĩa 5% là 0.182

Và ta thực hiện biến đổi dữ liệu với giá trị lambda vừa tìm được.



Hình 2.77: Phân phối trước và sau khi biến đổi của AggregateFollowers.

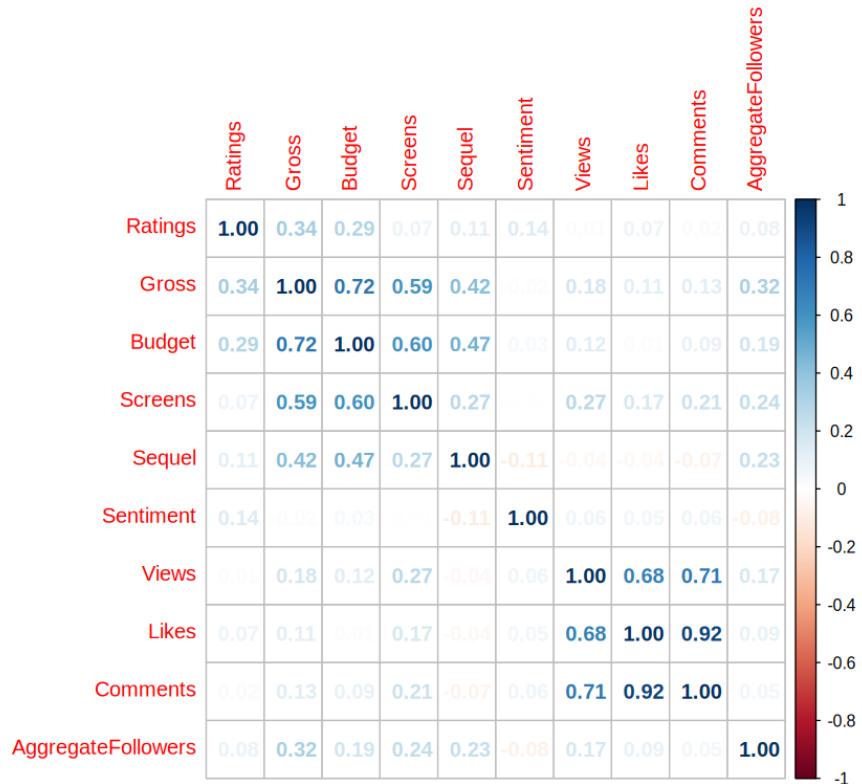
Nhận xét:

- Ta có được giá trị lambda tối ưu là 0.33 và sử dụng giá trị này để biến đổi biến AggregateFollowers. Biểu đồ histogram phía bên dưới thể hiện phân phối của biến này trước và sau khi biến đổi. Dễ dàng thấy được, sau khi biến đổi, biến này đã tương đối chuẩn hơn.

2.2.6. Mô hình hóa

Xử lý dữ liệu tuyến

Đầu tiên, ta trực quan hóa ma trận tương quan giữa các biến trong tập dữ liệu



Hình 2.78: Ma trận tương quan giữa các biến trong tập dữ liệu phim truyền thông.

Nhận xét:

- Một số cặp có khả năng đa cộng tuyến cao: Views' và 'Comments', Likes' và 'Comments', 'Gross' và 'Budget'
- Một số cặp tương quan nghịch tương đối vừa: Sequel và Sentiment (-0.11)

Ta xử lý đa cộng tuyến thông qua các bước sau. Bước 1: Tính toán chỉ số VIF

1	Ratings	Budget	Screens
2	Sequel		
3	1.196619	2.239092	1.740774
4		1.406844	
5	Sentiment	Views	Likes
6	Comments		
7	1.050186	2.172907	7.324670
8		7.884851	
9	AggregateFollowers		
10	1.147767		

Nhận xét:

- Ta có chọn ngưỡng bằng 3

Bước 2: Loại bỏ các biến dựa trên VIF nếu vượt quá ngưỡng

```
1      Ratings          Budget          Screens
2      Sequel
3      Sentiment         Views          Likes
4      AggregateFollowers
5
6 Call :
7 lm(formula = Gross ~ Ratings + Budget + Screens + Sequel +
8   Sentiment +
9   Views + Likes + AggregateFollowers, data = csm_df)
10
11 Residuals:
12      Min       1Q    Median      3Q      Max
13 -125889542 -27197502 -3079763 16371993 417347246
14
15 Coefficients:
16                               Estimate Std. Error t value Pr(>|t|)
17 (Intercept)             -1.226e+08 2.671e+07 -4.588 7.50e-06 ***
18 Ratings                  1.590e+07 4.005e+06  3.970 9.73e-05 ***
19 Budget                   7.509e-01 9.803e-02  7.660 5.68e-13 ***
20 Screens                  1.448e+04 3.372e+03  4.294 2.62e-05 ***
21 Sequel                   8.854e+06 4.451e+06  1.989 0.04788 *
22 Sentiment                -4.850e+05 5.402e+05 -0.898 0.37022
23 Views                    4.518e-01 1.158e+00  0.390 0.69692
24 Likes                     9.093e+01 1.766e+02  0.515 0.60717
25 AggregateFollowers     2.494e+00 8.657e-01  2.880 0.00436 **
26 ---
27 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
28
29 Residual standard error: 55930000 on 222 degrees of freedom
Multiple R-squared:  0.6179 ,    Adjusted R-squared:  0.6042
```

```
30 F-statistic: 44.88 on 8 and 222 DF, p-value: < 2.2e-16
```

Như vậy, chúng ta đã thu được tập dữ liệu bao gồm các biến không có đa cộng tuyến với nhau. (Đã loại bỏ được biến comments)

Khảo sát ngoại lai

Để loại bỏ những ngoại lai, ta sử dụng IQR để tìm và loại bỏ các điểm cực ngoại lai.

```
1 # Ở đây sử dụng IQR để loại bỏ ngoại lai
2 Q1 <- quantile(cleaned_df$'Gross', 0.25)
3 Q3 <- quantile(cleaned_df$'Gross', 0.75)
4 IQR <- Q3 - Q1
5 lower_bound <- Q1 - 1.5 * IQR
6 upper_bound <- Q3 + 1.5 * IQR
7 outliers <- which(cleaned_df$'Gross' < lower_bound | cleaned_df
   $'Gross' > upper_bound)
8 print(outliers)
9
10 lower_bound <- Q1 - 3 * IQR
11 upper_bound <- Q3 + 3 * IQR
12 extreme_outliers <- which(cleaned_df$'Gross' < lower_bound |
   cleaned_df$'Gross' > upper_bound)
13 print(extreme_outliers)
```

Ta thu được các điểm dữ liệu ngoại lai: 11 19 27 28 47 71 125 128 134 151 162 164 165 166 167 168

Các điểm dữ liệu cực ngoại lai: 11 47 128 164 165 166 167

Bằng cách loại bỏ các điểm dữ liệu cực ngoại lai, ta thu được tập dữ liệu để có thể khảo sát tiếp.

Xây dựng mô hình đầy đủ

Đầu tiên, ta phân chia tập dữ liệu thành 2 phần: train (80%) và test (20%).

```
1 split_ratio <- 0.8
2 split_index <- floor(nrow(csm_df) * split_ratio)
3
4 train = csm_df[1:split_index,]
5 test = csm_df[(split_index + 1):nrow(csm_df),]
```

Sau đó, ta xây dựng mô hình đầy đủ các biến như sau:

```
1 full_lm <- lm(Gross ~ ., data = train)
```

```
2 print(summary(full.lm))
```

Kết quả

```
1 lm(formula = Gross ~ ., data = train)
2
3 Residuals:
4   Min     1Q Median     3Q    Max
5 -136.99 -18.79    3.70   22.97   84.15
6
7 Coefficients:
8
9             Estimate Std. Error t value Pr(>|t|)
10 (Intercept) -60.21332  62.87576 -0.958 0.3396
11 Ratings      21.80737  12.55030  1.738 0.0841 .
12 Budget        0.65118  0.08619  7.555 2.46e-12 ***
13 Screens       2.89668  0.49989  5.795 3.25e-08 ***
14 Sequel         11.31840  6.07145  1.864 0.0640 .
15 Sentiment     -0.21697 10.96256 -0.020 0.9842
16 Views          -0.29107  0.25634 -1.136 0.2578
17 Likes          1.79341  0.75762  2.367 0.0190 *
18 AggregateFollowers 0.03477  0.09788  0.355 0.7228
19 ---
20 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
21
22 Residual standard error: 34.9 on 170 degrees of freedom
23 Multiple R-squared:  0.6614, Adjusted R-squared:  0.6455
24 F-statistic: 41.51 on 8 and 170 DF, p-value: < 2.2e-16
```

Nhận xét:

- Mô hình này có R-squared hiệu chỉnh 0.6455, tức là phương sai biến Gross có thể được giải thích được 64.55% dựa trên các biến độc lập của mô hình.
- Với mức ý nghĩa 5%, ta thấy các biến Budget, Screen, Ratings, Sequel và Likes có ý nghĩa đối với mô hình

Lựa chọn model tốt nhất

Với số lượng lớn các yếu tố dự đoán, điều quan trọng là phải giảm thiểu mô hình bằng cách chỉ bao gồm các yếu tố dự đoán hữu ích. Có tất cả 8 yếu tố dự đoán trong tập dữ liệu, nghĩ là có thể có 2^8 mô hình hồi quy. Để chọn mô hình một cách hiệu quả, việc lựa chọn lùi được thực hiện bằng

sử dụng step function. Phương pháp này lặp lại các quy trình để giảm thiểu Akaike's Information Criteria (AIC) và Bayesian Information Criteria (BIC). Lựa chọn mô hình ngược so với lựa chọn tiến vì nó loại bỏ khả năng một yếu tố dự đoán mới được chọn có khả năng tương tự hoặc nhiều hơn để giải thích các phần của phản hồi đã được giải thích bởi một yếu tố dự đoán khác có trong mô hình.

```

1 # Mô hình chặn dưới
2 model.lb <- lm(Gross ~ 1, data = train)
3
4 # Mô hình chặn trên
5 model.up <- full.lm
6
7 step(full.lm, scope = list(lower = model.lb, upper = model.up),
      direction = "both", trace = FALSE)

```

Kết quả

```

1 Call:
2 lm(formula = Gross ~ Ratings + Budget + Screens + Sequel +
     Likes,
     data = train)
3
4 Coefficients:
5 (Intercept)      Ratings       Budget       Screens       Sequel
6                         Likes
7 -72.5562        24.4101       0.6444        2.9817       11.9940
8                         1.0309

```

```

1 csm_models<- regsubsets(Gross ~ Ratings + Dislikes +
    AggregateFollowers, data = train)
2 summary.csm<-summary(csm_models)
3
4 # Lựa chọn mô hình tốt nhất từ reg subsets
5 summary.csm$which

```

Xây dựng mô hình tốt nhất dựa trên BIC

```

1 # Tiêu chí chọn mô hình tốt nhất 4: mô hình với BIC nhỏ
2 summary.csm$bic
3
4 best_model_index <- which.min(summary.csm$bic)
5 best_model <- summary.csm$which[best_model_index, ]

```

```

6 print(best_model)
7 best_vars <- names(best_model[best_model])
8 best_vars <- best_vars[best_vars != "(Intercept)"]
9 print(best_vars)
10
11 # Xây dựng mô hình tốt nhất
12 formula_str <- paste("Gross ~", paste(best_vars, collapse = " +
13 best_model_csm <- lm(as.formula(formula_str), data=train)
14
15 # Tóm tắt mô hình
16 summary(best_model_csm)

```

Kết quả:

```

1 lm(formula = as.formula(formula_str), data = train)
2
3 Residuals:
4      Min       1Q   Median       3Q      Max
5 -142.604   -17.715    1.929   22.419   87.326
6
7 Coefficients:
8                 Estimate Std. Error t value Pr(>|t|)
9 (Intercept) -27.97801   10.72422  -2.609 0.009870 ***
10 Budget       0.75150    0.07662   9.808 < 2e-16 ***
11 Screens      2.78579    0.48627   5.729 4.33e-08 ***
12 Likes        1.02351    0.29564   3.462 0.000674 ***
13 ---
14 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
15
16 Residual standard error: 35.3 on 175 degrees of freedom
17 Multiple R-squared:  0.6434 , Adjusted R-squared:  0.6373
18 F-statistic: 105.3 on 3 and 175 DF,  p-value: < 2.2e-16

```

Như vậy, ta thu được mô hình

$$\text{Gross} = -27.97801 * (\text{Intercept}) + 0.75150 * \text{Budget} + 2.78579 * \text{Screens} * 1.02351 * \text{Likes} \quad (2.1)$$

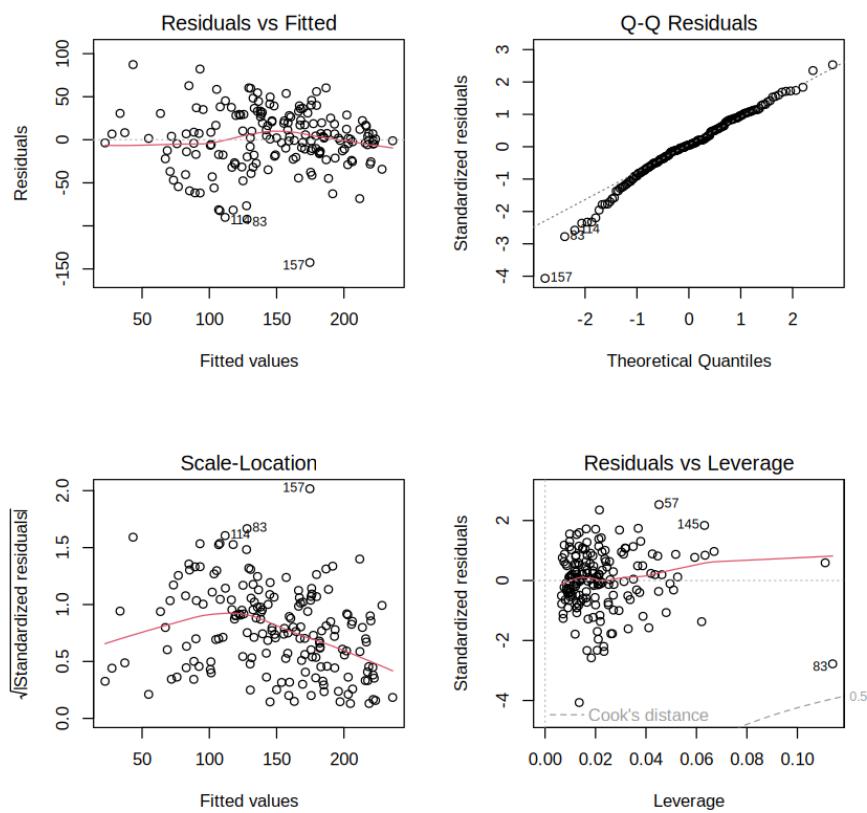
Nhận xét:

- R-squared hiệu chỉnh của mô hình là 0.6373, tức là 63.73% phương sai của biến Gross được giải

thích bởi các biến độc lập của mô hình được chọn.

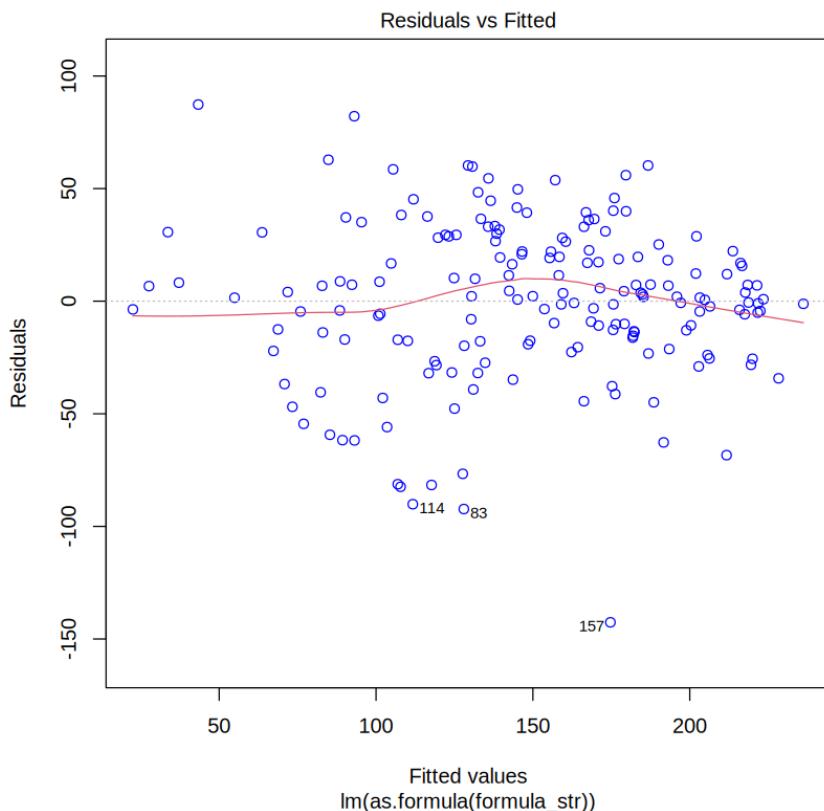
- Điều này có ý nghĩa là, biến Gross sẽ được giải thích thông qua hai biến Budget, Screens và Likes
- Số lượng rạp chiếu có ảnh hưởng nhiều đến với tổng doanh thu của một bộ phim. Số lượng rạp chiếu càng nhiều thì doanh thu của một bộ phim càng cao.
- Số lượt thích của một bộ phim cũng ảnh hưởng đến tổng doanh thu của một bộ phim. Một bộ phim được nhiều người yêu thích thì doanh thu càng cao.

Bây giờ, ta sẽ đi phân tích mô hình này.



Hình 2.79: .

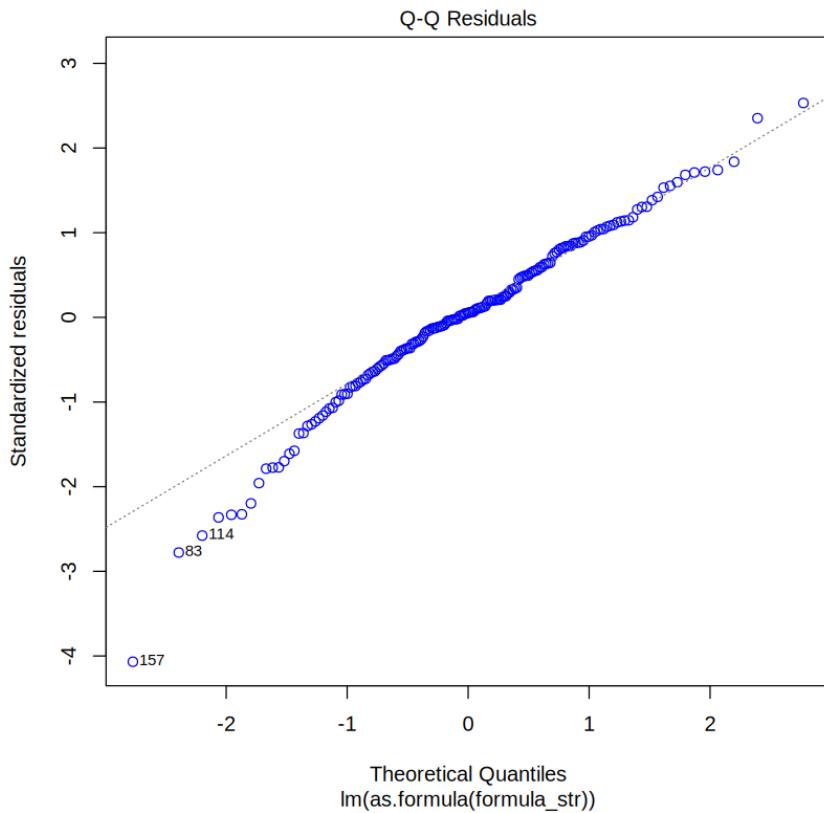
Phân tích Residuals vs Fitted Plot: Biểu đồ Residuals vs Fitted Plot đưa ra dấu hiệu nếu có các mẫu phi tuyến tính. Để hồi quy tuyến tính chính xác, dữ liệu cần phải tuyến tính nên điều này sẽ kiểm tra xem điều kiện đó có được đáp ứng hay không.



Hình 2.80: Biểu đồ Residuals vs Fitted Plot.

Dựa trên biểu đồ này, ta thấy đường cong màu đỏ có dáng gần như một đường thẳng, và các phần tử trái dọc theo đường cong này một cách tương đồng đều. Điều này chứng tỏ không có quan hệ phi tuyến xuất hiện trong dữ liệu.

Phân tích Normal Q–Q (quantile-quantile) Plot: Các giá trị thặng dư (residual) nên có phân phối chuẩn. Để kiểm tra điều này, chúng ta cần quan sát biểu đồ QQ Residuals plot, nếu các điểm được xếp thành một đường thẳng (hoặc gần như thẳng) thì chứng tỏ các giá trị thặng dư (residual) có phân phối chuẩn.



Hình 2.81: Normal Q–Q (quantile-quantile) Plot.

Như hình vẽ kết quả ở trên, ta thấy rõ điều đó, residual có phân phối chuẩn.

Cẩn thận hơn, chúng ta thử dùng Shapiro-Wilk test để kiểm tra có đúng thật là các giá trị thặng dư có phân phối chuẩn hay không?

- H0: Biến thặng dư của mô hình phân phối chuẩn trong một số quần thể.
- H1: Biến thặng dư của mô hình không phân phối chuẩn trong một số quần thể.

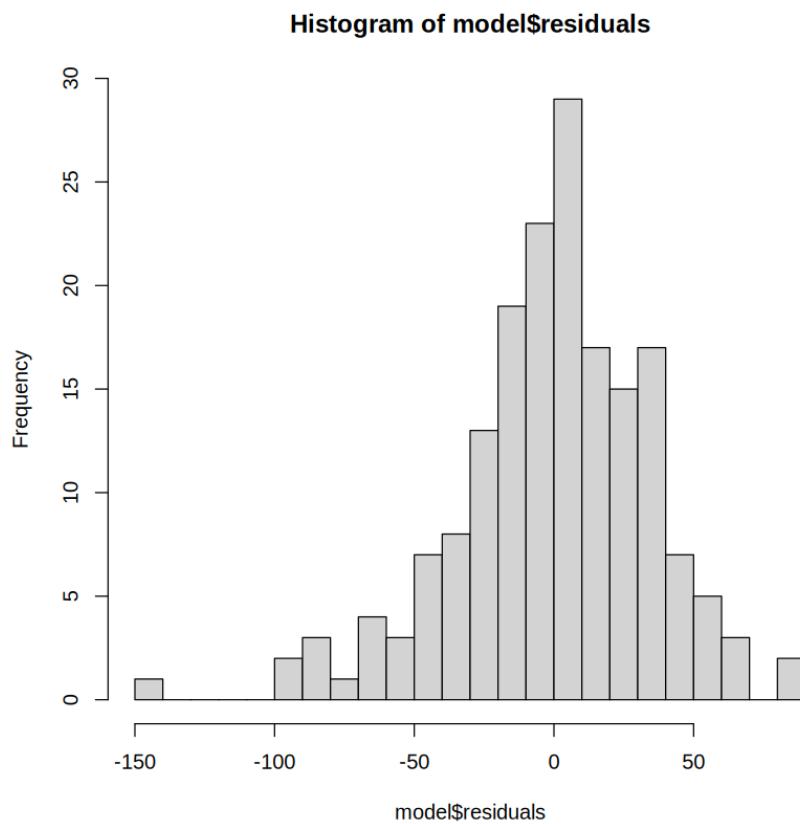
Kết quả:

```

1 Shapiro-Wilk normality test
2
3 data: model$residuals
4 W = 0.97565, p-value = 0.003158
5
6 [1] "H0 rejected: the residuals are NOT distributed normally"

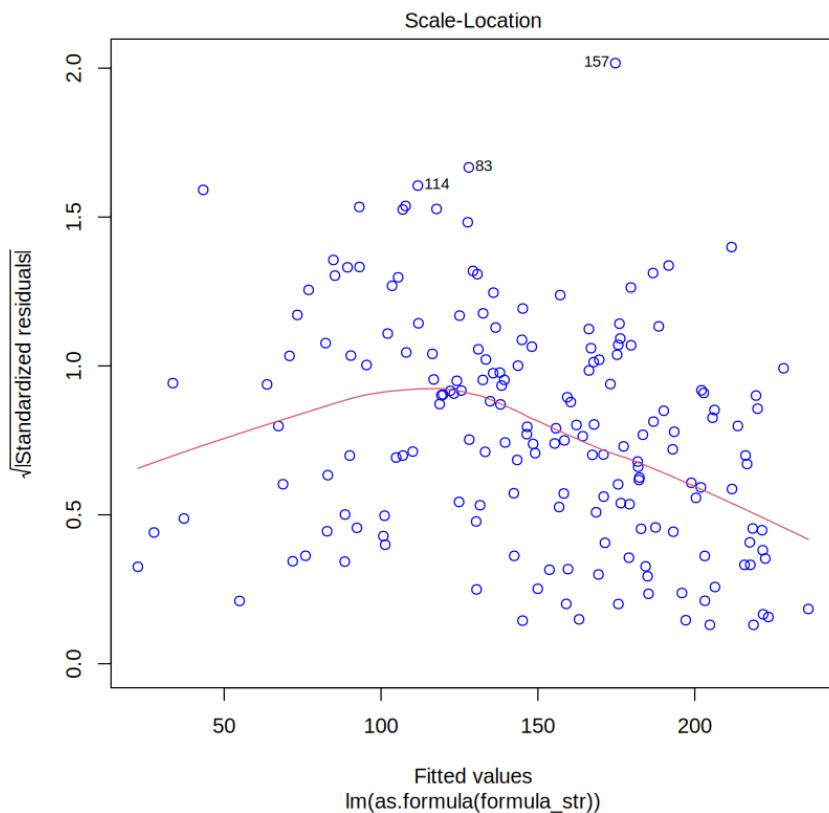
```

Kết quả cho thấy p-value bé hơn mức ý nghĩa alpha 0.05 nên ta có thể bác bỏ giả thuyết H0, biến thặng dư của chúng ta không chuẩn trong một số quần thể.



Hình 2.82: Histogram biến thặng dư của mô hình hồi quy CSM.

Phân tích Scale-Location: Biểu đồ scale-location kiểm định giả định hồi quy về phương sai bằng nhau (homoscedasticity), tức là giá trị thặng dư có phương sai bằng với đường hồi quy.



Hình 2.83: Scale-Location Plot.

Nhận xét:

- Đường màu đỏ gần bị lệch về phía dưới của biểu đồ. Nghĩa là, độ phân tán của giá trị thặng dư gần không bằng nhau ở tất cả các giá trị phù hợp.
- Các giá trị thặng dư được phân tán ngẫu nhiên xung quanh đường màu đỏ với độ biến thiên tương đối bằng nhau ở tất cả các giá trị phù hợp.

Cẩn thận hơn, chúng ta sử dụng Breusch-Pagan test để kiểm tra có thật là như vậy không?

- H0: Các giá trị thặng dư là homoscedastic
- H1: Các giá trị thặng dư là heteroscedastic

Kết quả:

```

1 studentized Breusch-Pagan test
2
3 data: model
4 BP = 6.8056, df = 3, p-value = 0.07836

```

```

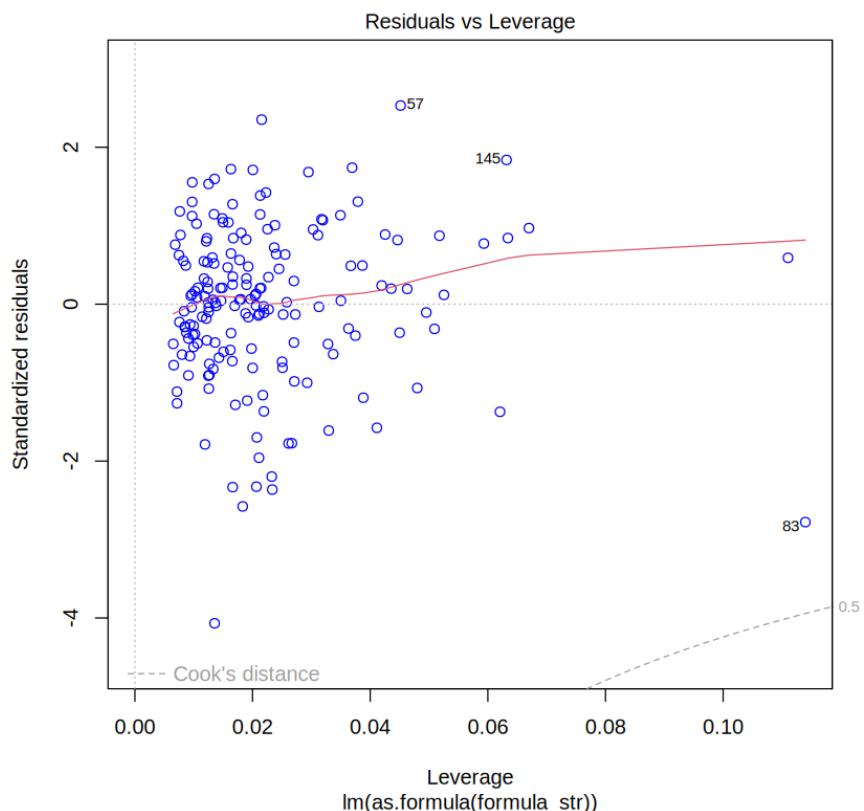
5
6 [1] "H0 failed to reject: Error variance spreads CONSTANTLY (
    Homoscedasticity)"

```

Như vậy, ta thấy p-value lớn hơn mức ý nghĩa 0.05, ta chưa đủ điều kiện bác bỏ H0. Vậy các giá trị thăng dư là homoscedastic

Phân tích Residuals vs Leverage Biểu đồ này có thể được sử dụng để tìm các trường hợp có ảnh hưởng trong tập dữ liệu. Một trường hợp có ảnh hưởng là một trường hợp mà nếu bị loại bỏ sẽ ảnh hưởng đến mô hình nên việc đưa vào hoặc loại trừ nó cần được xem xét.

Một trường hợp có ảnh hưởng có thể là một trường hợp ngoại lệ hoặc không và mục đích của biểu đồ này là xác định các trường hợp có ảnh hưởng lớn đến mô hình. Các ngoại lệ sẽ có xu hướng có giá trị cực cao hoặc cực thấp và do đó ảnh hưởng đến mô hình.

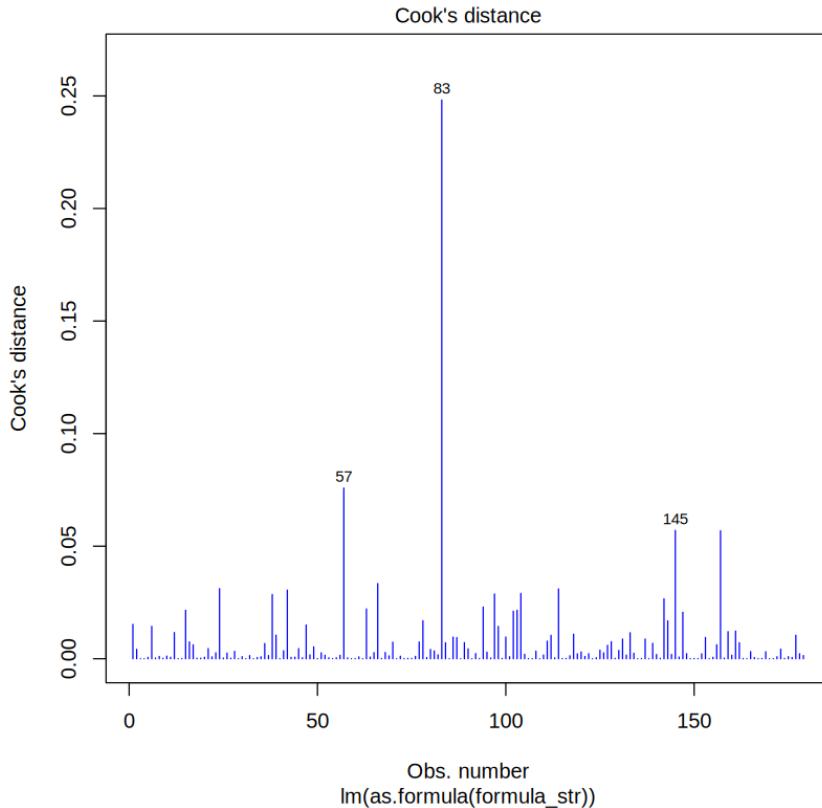


Hình 2.84: Residuals vs Leverage Plot.

Nhận xét:

- Một số điểm như 57, 83, 145 có thể là các điểm ngoại lai. Ta có thể thử loại bỏ để nâng cao chất lượng mô hình.

Ta cũng có thể nhận thấy có một số giá trị ngoại lai ở cách xa đường thẳng giữa. Ta có thể xem rõ hơn thông qua histogram của Cook's Distance



Hình 2.85: Cook Distance Plot của mô hình hồi quy CSM.

Loại bỏ ngoại lai dựa trên Cook's Distance

Sau khi loại bỏ những điểm có tiềm năng là ngoại lai, ta thu được các mô hình. Dựa trên các tiêu chí và R-squared hiệu chỉnh, đánh giá kiểm định phần thặng dư và đồng nhất phương sai. Ta vẫn lựa chọn được mô hình khi loại bỏ vừa đủ các điểm tiềm năng (Xem file code).

```

1 lm(formula = as.formula(formula_str), data = train[-c(
2   influential_points),
3   ])
4
4 Residuals:
5   Min     1Q Median     3Q    Max
6 -71.07 -17.18    0.42  19.32  58.85
7
8 Coefficients:

```

```

9             Estimate Std. Error t value Pr(>|t|)
10 (Intercept) -21.37468   9.25407  -2.310   0.0222 *
11 Budget       0.65659   0.06692   9.811 < 2e-16 ***
12 Screens      3.27610   0.41667   7.863 4.99e-13 ***
13 Likes        1.14767   0.25277   4.540 1.09e-05 ***
14 ---
15 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
16
17 Residual standard error: 27.76 on 162 degrees of freedom
18 Multiple R-squared:  0.7377 , Adjusted R-squared:  0.7329
19 F-statistic: 151.9 on 3 and 162 DF, p-value: < 2.2e-16

```

Nhận xét:

- R-squared hiệu chỉnh của mô hình là 0.7329, có nghĩa là 73.29% phương sai của Gross có thể được giải thích bởi các biến độc lập của mô hình.

Kiểm định phân phối chuẩn cho biến thặng dư. Ta có kết quả như sau:

```

1 Shapiro-Wilk normality test
2
3 data: model$residuals
4 W = 0.98973, p-value = 0.2725
5
6 [1] "H0 failed to reject: the residuals ARE distributed
     normally"

```

Kết quả cho thấy p-value lớn hơn mức ý nghĩa alpha 0.05 nên ta có thể bác bỏ giả thuyết H0, biến thặng dư của chúng ta chuẩn trong một số quản thể.

Và kiểm định đồng nhất phương sai. Ta có kết quả:

```

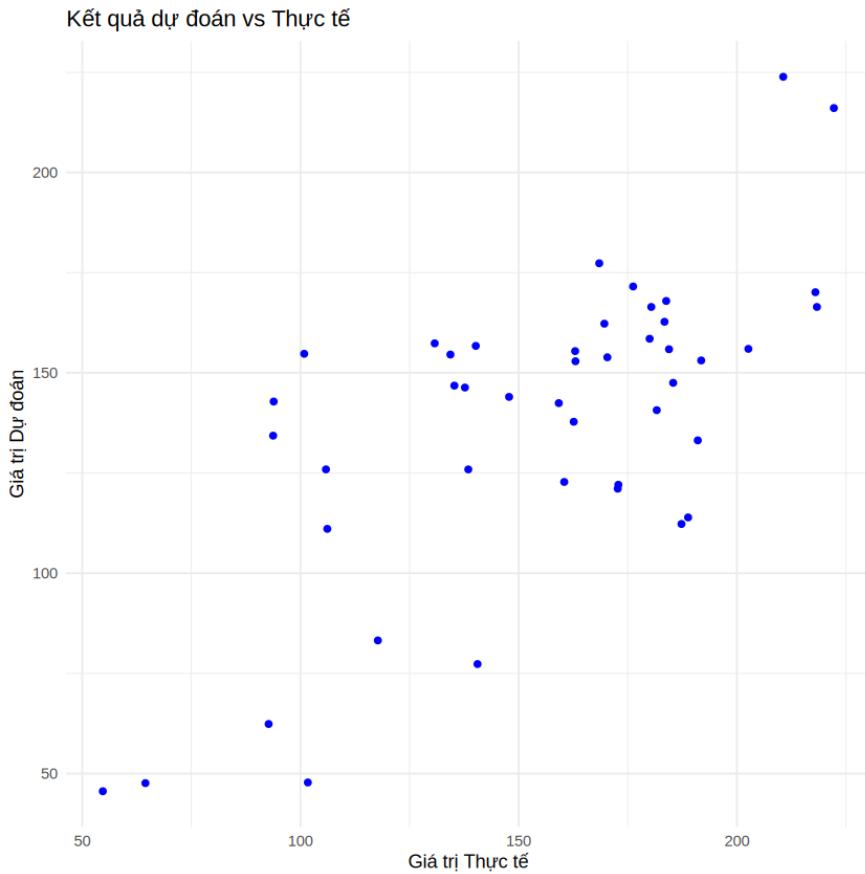
1 studentized Breusch-Pagan test
2
3 data: model
4 BP = 11.356, df = 3, p-value = 0.009948
5
6 [1] "H0 rejected: Error variance spreads INCONSTANTLY/
     generating patterns (Heteroscedasticity)"

```

Như vậy, ta thấy p-value nhỏ hơn mức ý nghĩa 0.05, ta chưa đủ điều kiện bác bỏ H0. Vậy các giá trị thặng dư là heteroscedasticity

Dự đoán và đánh giá kết quả

Ta trực quan kết quả dự đoán và đánh giá RMSE của mô hình đã chọn.



Hình 2.86: Trực quan kết quả dự đoán của mô hình tốt nhất. RMSE = 196.83.

Hiệu suất trên các độ đo:

- "MSE: 1231.319394"
- "RMSE: 35.090161"
- "MAE: 28.99513"
- "Correlation: 0.694378"
- " $\hat{R^2}$ between y_pred & y_true: 0.482161"

2.2.7. Mô hình hóa bằng PCR

Principal Component Regression (PCR) là một kỹ thuật kết hợp Phân tích thành phần chính (PCA) và hồi quy tuyến tính để giải quyết đa cộng tuyến và giảm chiều trong các tập dữ liệu cao chiều. Các bước chính trong PCR là:

- PCA biến đổi các biến dự báo ban đầu thành một tập hợp các biến mới, không tương quan được gọi là các thành phần chính. Các thành phần này là các tổ hợp tuyến tính của các biến ban đầu và được sắp xếp theo lượng phương sai mà chúng giải thích trong dữ liệu. Mỗi thành phần chính nắm bắt được phương sai tối đa có thể trong khi vẫn trực giao với các thành phần trước đó.
- Một tập hợp con các thành phần chính (giải thích phương sai lớn nhất) được chọn và sử dụng làm các yếu tố dự báo trong mô hình hồi quy tuyến tính để dự báo biến phản hồi. Bằng cách tập trung vào các thành phần chính nắm bắt được phương sai lớn nhất, PCR hướng đến mục tiêu xây dựng một mô hình hồi quy ổn định và dễ diễn giải hơn.

```
1 pcr_model <- pcr(`Gross` ~ ., data = train, scale = TRUE,
  validation = "CV") # Fit PCR model with cross-validation
```

Đối số xác thực = “CV” chỉ định rằng xác thực chéo (cross-validation - CV) nên được sử dụng để xác thực mô hình. Xác thực chéo là một phương pháp mạnh mẽ để đánh giá hiệu suất dự đoán của một mô hình. Nó bao gồm việc phân vùng dữ liệu thành các tập hợp con, huấn luyện mô hình trên một số tập hợp con (bộ huấn luyện - training set) và xác thực nó trên các tập hợp con còn lại (bộ xác thực - validation set). Quá trình này được lặp lại nhiều lần để đảm bảo hiệu suất của mô hình là nhất quán và không phụ thuộc vào phân vùng dữ liệu cụ thể.

Bằng cách sử dụng xác thực chéo, mô hình ít có khả năng quá khớp với dữ liệu huấn luyện. Quá khớp xảy ra khi mô hình nắm bắt được nhiều và các mẫu cụ thể trong dữ liệu huấn luyện không tổng quát hóa thành dữ liệu mới, chưa từng thấy. Xác thực chéo giúp phát hiện và giảm thiểu tình trạng quá khớp bằng cách kiểm tra mô hình trên các tập hợp con khác nhau của dữ liệu.

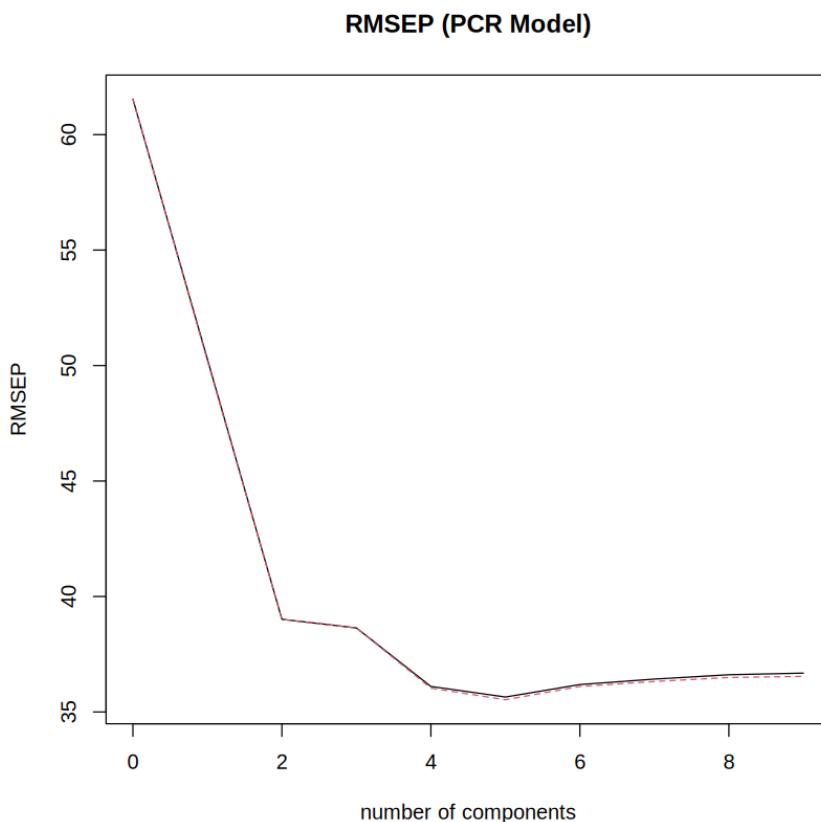
```
1 Data:   X dimension: 184 9
2                 Y dimension: 184 1
3 Fit method: svdpc
4 Number of components considered: 9
5
6 VALIDATION: RMSEP
7 Cross-validated using 10 random segments.
8             (Intercept) 1 comps  2 comps  3 comps  4 comps  5 comps
9                 CV       61.54    50.28    39.01    38.64    36.10    35.65
10                adjCV   36.19
11
12 adjCV       CV       61.54    50.21    39.01    38.64    36.03    35.53
13                adjCV  36.10
14
15 adjCV       CV       36.42    36.61    36.68
16                adjCV  36.32    36.49    36.54
```

```

14
15 TRAINING: % variance explained
16      1 comps   2 comps   3 comps   4 comps   5 comps   6 comps   7
17          comps   8 comps
18 X           36.99     56.35    69.36    79.30    87.38    94.55
19         98.22     99.47
20 Gross       34.19     61.29    62.20    67.11    68.00    68.02
21         68.39     68.44
22         9 comps
23 X           100.00
24 Gross       69.01

```

Lựa chọn số lượng thành phần chính: Để quyết định được số thành phần chính tối ưu, chúng ta cần phải trung hòa giữa độ phức tạp của mô hình (tức là số lượng components) và RMSEP (Root Mean Squared Error of Prediction).



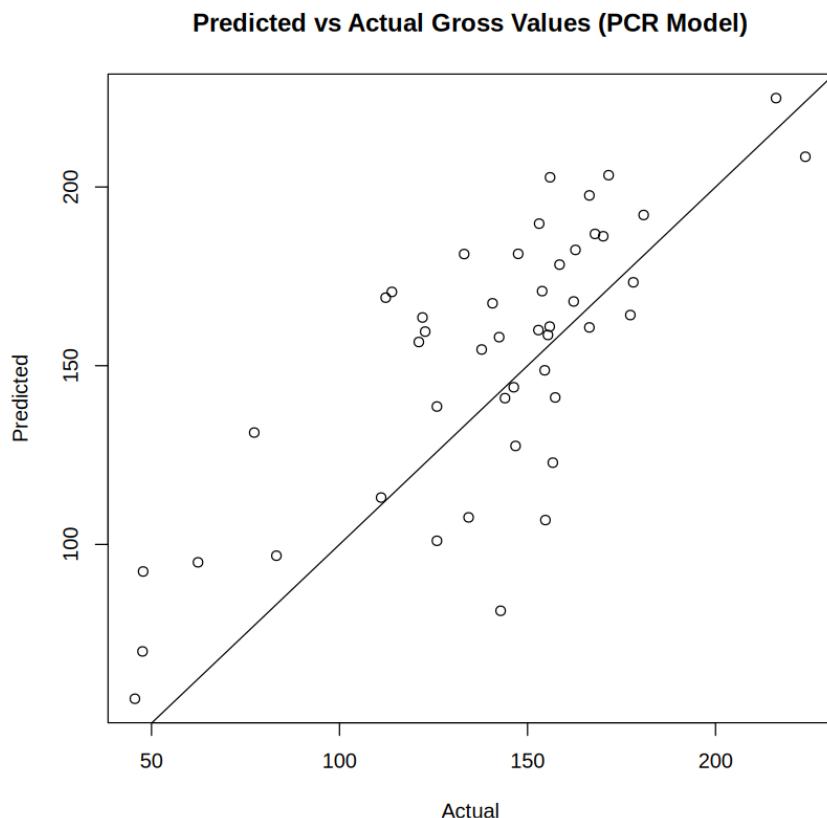
Hình 2.87: Giá trị RMSEP với số lượng thành phần chính khác nhau.

- RMSEP Values: 5 components - 35.53 (nhỏ nhất)

- Variance Explained: 5 components - 68.00%

Do đó ta chọn 5 components.

Đánh giá kết quả dự đoán



Hình 2.88: Kết quả dự đoán của mô hình PCR.

```

1 # Calculate and print the Root Mean Squared Error (RMSE)
2 rmse <- sqrt(mean((test$`Gross` - predictions)^2)) # Calculate
   RMSE between actual and predicted values
3 print(paste("RMSE: ", rmse))

```

Kết quả: RMSE: 29.0363826627182

```

1 # Calculate the sum of squares of residuals
2 ss_res <- sum((test$`Gross` - predictions)^2)
3
4 # Calculate the total sum of squares
5 ss_tot <- sum((test$`Gross` - mean(test$`Gross`))^2)
6

```

```

7 # Calculate R-squared
8 r_squared <- 1 - (ss_res / ss_tot)
9
10 # Print R-squared
11 print(paste("R-squared: ", r_squared))

```

Kết quả: 0.418260507745999

Giá trị R-squared 0.4182 có nghĩa là 41.82% phương sai của biến phụ thuộc Gross được giải thích bởi các biến độc lập của mô hình.

Kiểm định thặng dư và đồng nhất phương sai cho kết quả:

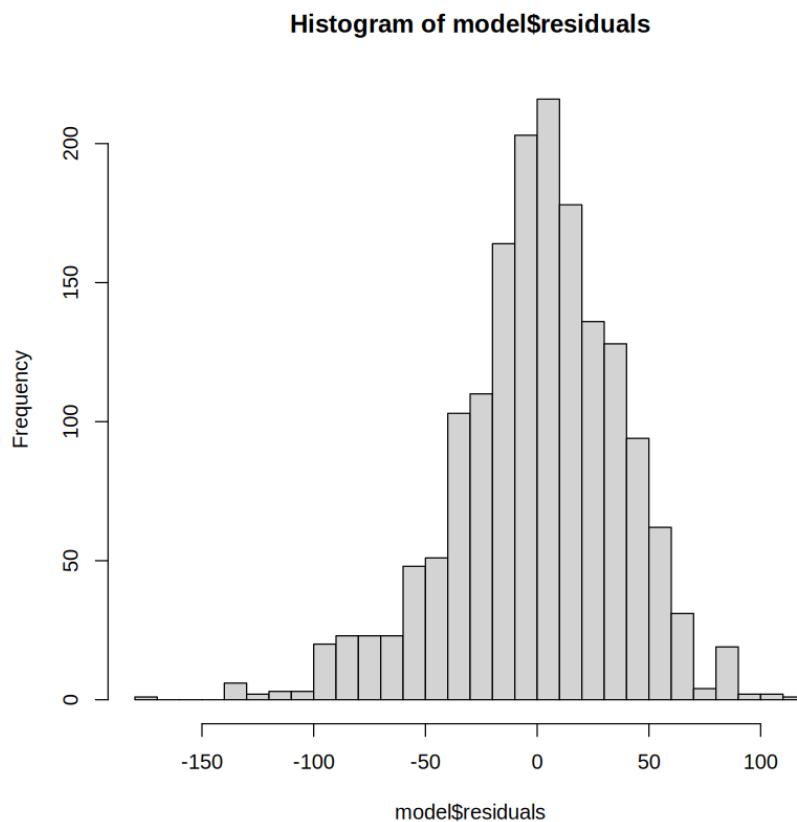
```

1 Shapiro-Wilk normality test
2
3 data: model$residuals
4 W = 0.98315, p-value = 4.99e-13
5
6 [1] "H0 rejected: the residuals are NOT distributed normally"

```

Nhận xét:

- Mô hình thỏa mãn giả định phân phối thặng dư không xấp xỉ chuẩn



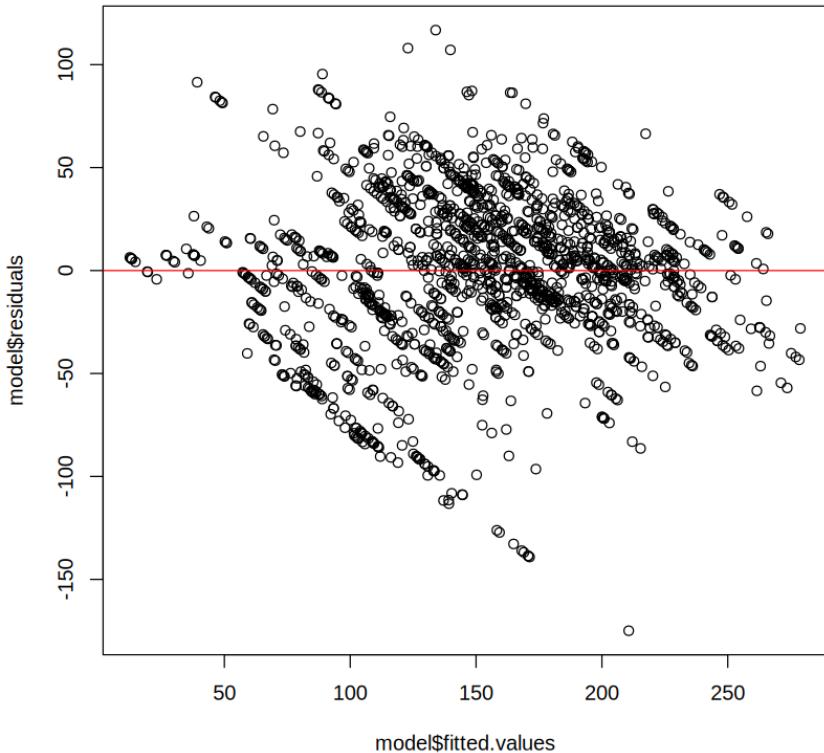
Hình 2.89: Kiểm định tính chuẩn thường dư của mô hình PCR.

```

1 studentized Breusch-Pagan test
2
3 data:  model
4 BP = 16.065, df = 9, p-value = 0.06554
5
6 [1] "H0 failed to reject: Error variance spreads CONSTANTLY (
    Homoscedasticity)"

```

- Mô hình không đồng nhất phương sai



Hình 2.90: Kiểm định đồng nhất phương sai của mô hình PCR.

2.2.8. Mô hình hóa bằng PLS

Partial Least Squares Regression là một kỹ thuật, không giống như PCR, xem xét cả các biến dự báo và biến phản hồi trong quá trình giảm chiều. Các bước chính trong PLS là:

- Latent Variable Extraction: PLS trích xuất một tập hợp các biến tiềm ẩn (thành phần) tối đa hóa hiệp phương sai giữa các biến dự báo và biến phản hồi. Các thành phần này là các tổ hợp tuyến tính của các biến ban đầu, được chọn theo cách mà chúng nắm bắt được càng nhiều thông tin có liên quan càng tốt để dự đoán biến phản hồi. Điều này đảm bảo rằng các thành phần được trích xuất có liên quan trực tiếp đến kết quả quan tâm.
- Regression: Các biến tiềm ẩn sau đó được sử dụng làm biến dự báo trong mô hình hồi quy tuyến tính để dự báo biến phản hồi. Bằng cách kết hợp biến phản hồi vào quy trình trích xuất thành phần, PLS hướng đến mục tiêu cải thiện độ chính xác dự báo của mô hình hồi quy.

```

1 Data:   X dimension: 184 9
2                 Y dimension: 184 1
3 Fit method: kernelpls

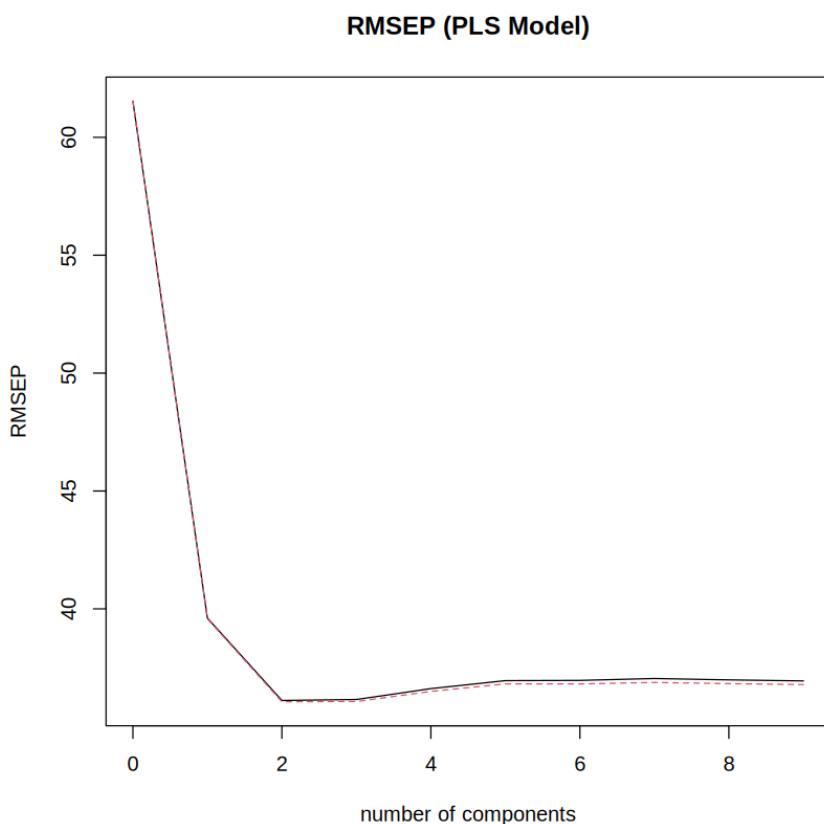
```

```

4 Number of components considered: 9
5
6 VALIDATION: RMSEP
7 Cross-validated using 10 random segments.
8     (Intercept) 1 comps 2 comps 3 comps 4 comps 5 comps
9             6 comps
9 CV          61.54    39.76    36.29    36.20    36.70    36.84
10            37.03
10 adjCV      61.54    39.71    36.22    36.12    36.58    36.71
11            36.88
11            7 comps 8 comps 9 comps
12 CV          37.20    37.15    37.11
13 adjCV      37.03    36.98    36.95
14
15 TRAINING: % variance explained
16     1 comps 2 comps 3 comps 4 comps 5 comps 6 comps 7
17             comps 8 comps
17 X          33.56    55.01    66.00    73.38    81.48    86.81
18            91.62    94.29
18 Gross      59.95    67.52    68.43    68.62    68.75    68.91
19            69.00    69.01
19            9 comps
20 X          100.00
21 Gross      69.01

```

Lựa chọn số lượng thành phần chính: Để quyết định được số thành phần chính tối ưu, chúng ta cần phải trung hòa giữa độ phức tạp của mô hình (tức là số lượng components) và RMSEP (Root Mean Squared Error of Prediction).

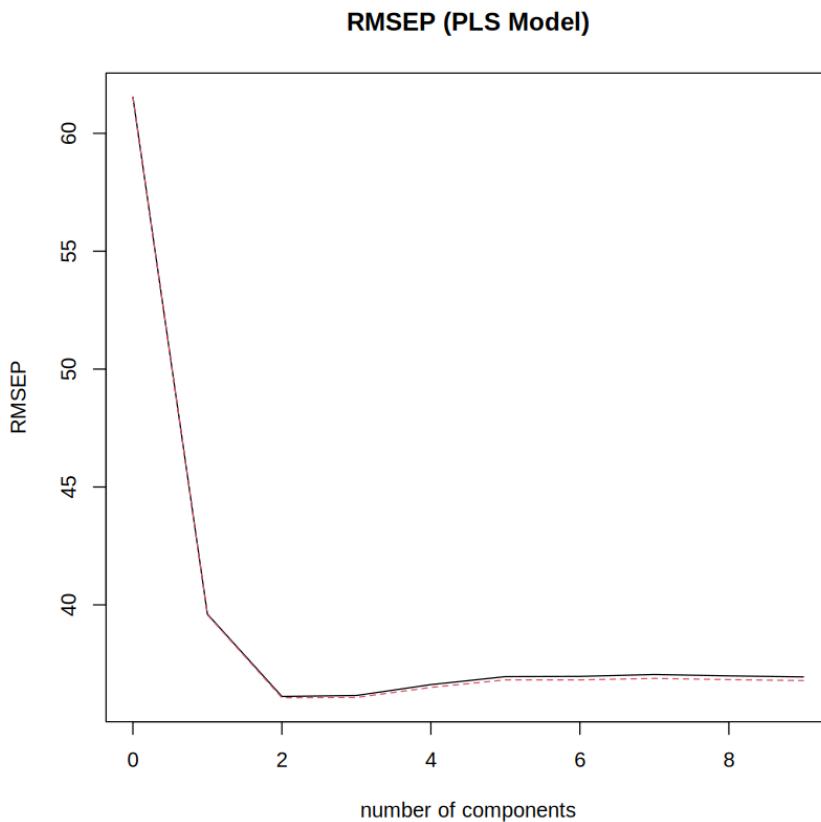


Hình 2.91: Giá trị RMSEP với số lượng thành phần chính khác nhau.

- RMSEP Values: 2 components - 36.22
- Variance Explained: 3 components - 67.52%

Do đó ta chọn 2 components.

Đánh giá kết quả dự đoán



Hình 2.92: Kết quả dự đoán của mô hình PLS.

```

1 # Calculate and print the Root Mean Squared Error (RMSE)
2 rmse <- sqrt(mean((test$`Gross` - predictions2)^2)) #
   Calculate RMSE between actual and predicted values
3 print(paste("RMSE: ", rmse))

```

Kết quả: RMSE: 31.276

```

1 # Calculate the sum of squares of residuals
2 ss_res <- sum((test$`Gross` - predictions2)^2)
3
4 # Calculate the total sum of squares
5 ss_tot <- sum((test$`Gross` - mean(test$`Gross`))^2)
6
7 # Calculate R-squared
8 r_squared <- 1 - (ss_res / ss_tot)
9
10 # Print R-squared

```

```
11 print(paste("R-squared: ", r_squared))
```

Kết quả: R-squared: 0.4259

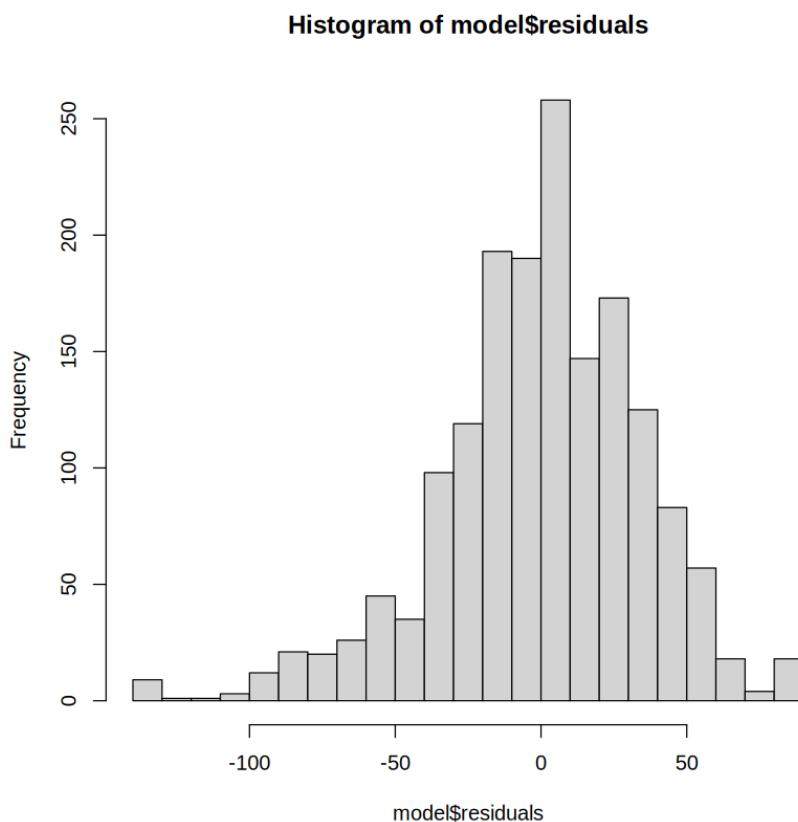
Giá trị R-squared 0.4259 có nghĩa là 42.59% phương sai của biến phụ thuộc ‘Gross’ được giải thích bởi các biến độc lập của mô hình.

Kiểm định thặng dư và đồng nhất phương sai cho kết quả:

```
1 Shapiro-Wilk normality test
2
3 data: model$residuals
4 W = 0.97724, p-value = 1.508e-15
5
6 [1] "H0 rejected: the residuals are NOT distributed normally"
```

Nhận xét:

- Mô hình thỏa mãn giả định phân phối thặng dư không xấp xỉ chuẩn



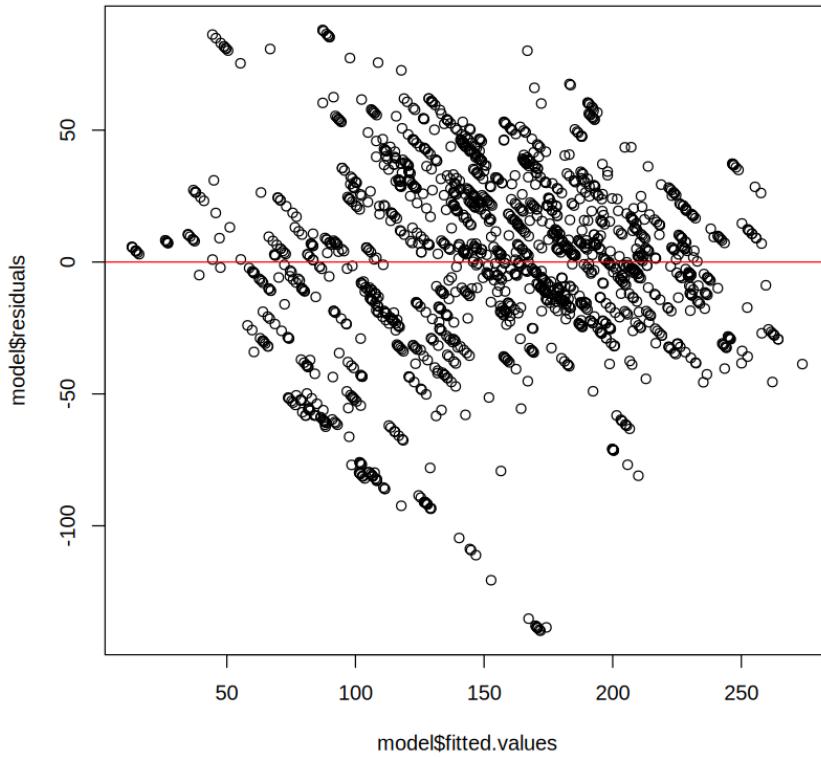
Hình 2.93: Kiểm định tính chuẩn thặng dư của mô hình PCR.

```

1 studentized Breusch-Pagan test
2
3 data: model
4 BP = 16.065, df = 9, p-value = 0.06554
5
6 [1] "H0 failed to reject: Error variance spreads CONSTANTLY (
    Homoscedasticity)"

```

- Mô hình không đồng nhất phương sai



Hình 2.94: Kiểm định đồng nhất phương sai của mô hình PCR.

2.2.9. So sánh hai mô hình PCR và PLS

So sánh về hiệu suất về mặt độ đo

- RMSEP:

- PCR RMSE: 29.0363

- PLS RMSE: 31.2765

- R-Squared:

- PCR: 0.4182
- PLS: 0.4259

Nhận xét:

- RMSE đánh giá độ lớn trung bình của các lỗi giữa giá trị dự đoán và giá trị thực tế. RMSE thấp hơn đáng kể của mô hình PLS cho thấy độ chính xác vượt trội của nó trong việc dự đoán doanh thu Gross, làm nổi bật hiệu quả của nó trong các ứng dụng thực tế khi mà các dự đoán chính xác là rất quan trọng.
- R-squared là tỷ lệ phương sai trong biến phụ thuộc có thể dự đoán được từ các biến độc lập. R-bình phương cao hơn của mô hình PLS cho thấy nó nắm bắt được mức độ biến động lớn hơn trong tổng doanh thu Gross, cho thấy sự phù hợp tổng thể tốt hơn với dữ liệu.

Phân tích độ phức tạp của mô hình

- PCR: Tập trung vào việc nắm bắt phương sai tối đa trong các yếu tố dự báo. Nó có hiệu quả trong việc giảm chiều và khám phá cấu trúc trong dữ liệu cao chiều. Tuy nhiên, hạn chế chính là các thành phần có phương sai cao nhất trong các yếu tố dự báo có thể không phải là thành phần có liên quan nhất để dự báo biến phản hồi.
- PLS: Nhắm mục đích tối đa hóa hiệp phương sai giữa các yếu tố dự báo và biến phản hồi. Phương pháp này không chỉ làm giảm tính cao chiều mà còn đảm bảo rằng các thành phần được giữ lại có liên quan trực tiếp đến kết quả quan tâm. Điểm mạnh của PLS nằm ở khả năng xác định các thành phần có khả năng dự báo phản hồi cao nhất, khiến nó rất phù hợp cho các nghiên cứu tập trung vào dự báo.

Phân tích điểm mạnh điểm yếu của PCR

- Điểm mạnh:
 - Tuyệt vời cho phân tích dữ liệu khám phá và hiểu cấu trúc cơ bản của dữ liệu.
 - Hiệu quả trong các tình huống mà việc hiểu cấu trúc phương sai trong các yếu tố dự báo quan trọng hơn việc dự đoán một kết quả cụ thể.
- Điểm yếu: Có thể bỏ qua mối quan hệ giữa các yếu tố dự báo và biến phản hồi. Không lý tưởng cho mô hình dự báo khi mục tiêu là giảm thiểu lỗi dự báo.

Phân tích điểm mạnh điểm yếu của PLS

- Trực tiếp nhắm mục tiêu vào phương sai có khả năng dự báo nhất của biến phản hồi, nâng cao độ chính xác dự báo.
- Rất hiệu quả trong việc xử lý dữ liệu đa cộng tuyến, khiến nó trở nên lý tưởng cho các tập dữ liệu phức tạp trong đó các yếu tố dự báo có mối quan hệ với nhau.

Khi nào nên dùng PCR

- Ưu tiên PCR khi mối quan tâm chính là giảm chiều và hiểu phương sai trong các yếu tố dự báo, không phụ thuộc vào tác động của chúng lên biến phản hồi.
- Phù hợp hơn với các phân tích thăm dò nhằm khám phá các cấu trúc ẩn trong dữ liệu, có thể không nhất thiết liên kết trực tiếp với biến mục tiêu.

Khi nào nên dùng PLS

- Chọn PLS khi cần độ chính xác dự đoán cao, đặc biệt là trong các tập dữ liệu phức tạp, trong đó các yếu tố dự đoán có tính cộng tuyến cao.
- Phù hợp khi tập trung vào việc nắm bắt và mô hình hóa các mối quan hệ cơ bản ảnh hưởng trực tiếp đến biến mục tiêu.

Trong tình huống tập trung vào khía cạnh dự đoán tổng doanh thu, chúng tôi ưu tiên sử dụng PLS.

CHƯƠNG 3

HOẠT ĐỘNG 2

3.1. Phân tích chất lượng rượu

3.1.1. Giới thiệu chung

Bộ dữ liệu về chất lượng rượu vang, có sẵn tại UCI Machine Learning Repository, chứa thông tin về các loại rượu vang đỏ và trắng từ vùng phía Bắc Bồ Đào Nha. Bộ dữ liệu này được sử dụng rộng rãi trong các nghiên cứu về học máy và phân tích dữ liệu nhằm dự đoán chất lượng của rượu dựa trên các đặc tính hóa học của nó.

- Bộ dữ liệu này được cung cấp bởi Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos và José Reis.
- Bao gồm hai tệp riêng biệt cho rượu vang đỏ và rượu vang trắng.
- Mỗi tệp chứa các giá trị về các thuộc tính hóa học và một cột chỉ số chất lượng (quality) từ 0 đến 10.

3.1.2. Phát biểu bài toán

Mục tiêu của đồ án này là giải quyết phương trình mô hình cuối cùng và xuất ra các giá trị thống kê như R-squared điều chỉnh, Mean Squared Error (MSE), Root Mean Squared Error (RMSE) và Mean Absolute Error (MAE). Đồng thời, đồ án sẽ kiểm tra mô hình bằng cách sử dụng các số liệu và hình ảnh minh họa để đánh giá tính tuyến tính của các tham số mô hình, kiểm tra tính độc lập tuần tự của các sai số, tính đồng nhất của phương sai (heteroscedasticity), tính bình thường của phân phối phần dư và đa cộng tuyến (multicollinearity). Ngoài ra, đồ án cũng sẽ xem xét các yếu tố khác như liệu có bất kỳ giá trị ngoại lệ (outliers) nào không và liệu có dữ liệu bị thiếu hay không. Cuối cùng, mô hình sẽ được kiểm tra bằng cách sử dụng bộ dữ liệu kiểm tra và kết quả sẽ được thảo luận chi tiết.

3.1.3. Phân tích chất lượng rượu trắng

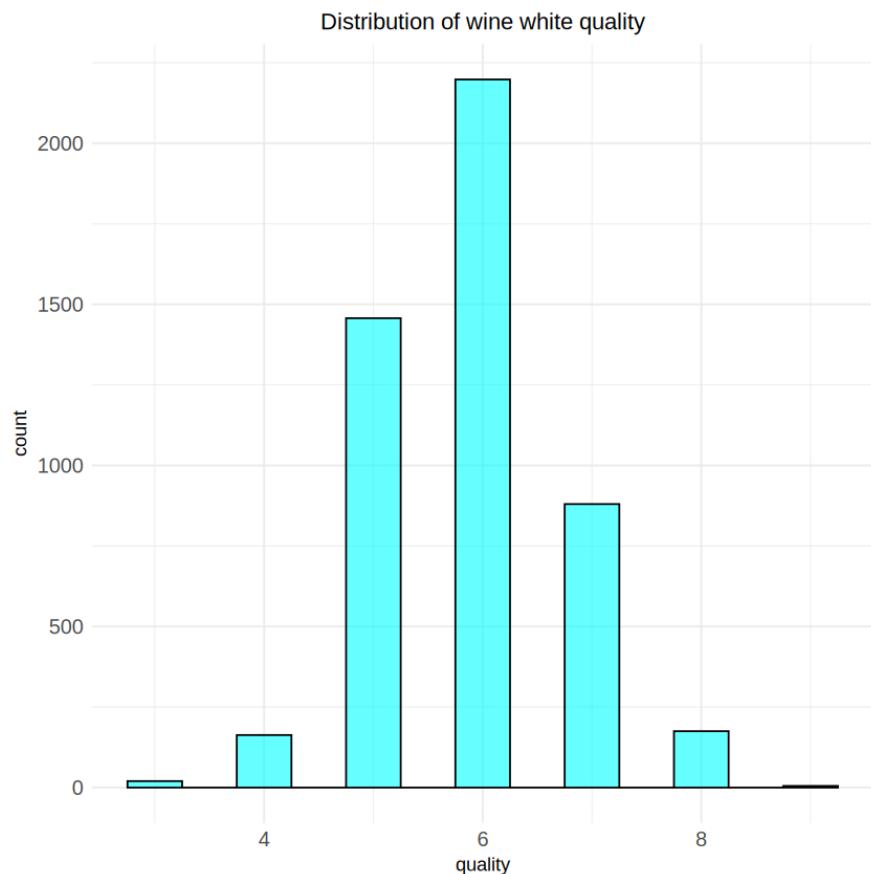
Các thông tin thống kê mô tả về bộ dữ liệu

variable	missing	min	lower	median	upper
max					

	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
2	>						
3	1 fixed.acidity	0	3.8	6.3	6.8	7.3	14.2
4	2 volatile.acidity	0	0.08	0.2	0.3	0.3	1.1
5	3 citric.acid	0	0	0.3	0.3	0.4	1.66
6	4 residual.sugar	0	0.6	1.7	5.2	9.9	65.8
7	5 chlorides	0	0.009	0	0	0	
		0.346					
8	6 free.sulfur.dioxide	0	2	23	34	46	289
9	7 total.sulfur.dioxide	0	9	108	134	167	440
10	8 density	0	0.987	1	1	1	1.04
11	9 pH	0	2.72	3.1	3.2	3.3	3.82
12	10 sulphates	0	0.22	0.4	0.5	0.6	1.08
13	11 alcohol	0	8	9.5	10.4	11.4	14.2
14	12 quality	0	3	5	6	6	9

Phân tích đơn biến

Chất lượng rượu

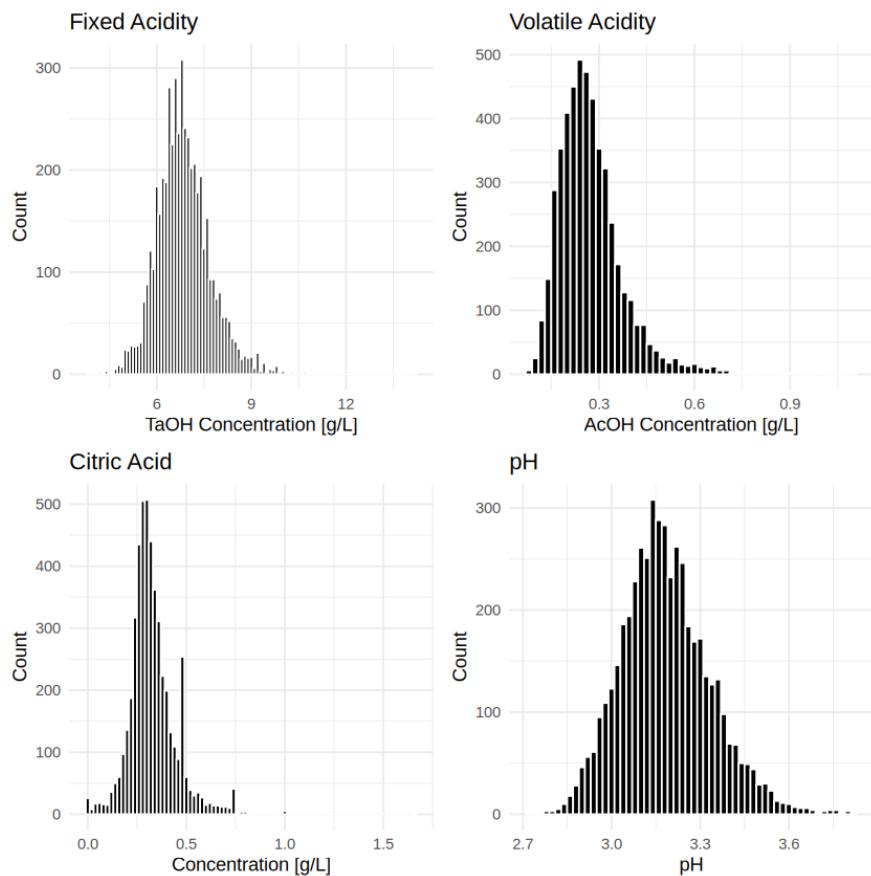


Hình 3.1: Chất lượng rượu trắng.

Nhận xét:

- Chất lượng rượu có phân phối đối xứng
- Hầu hết chất lượng rượu dở nằm ở mức 5, 6
- Không có rượu trắng nào đạt điểm tuyệt đối
- Chất lượng rượu trắng tệ nhất có điểm số là 3

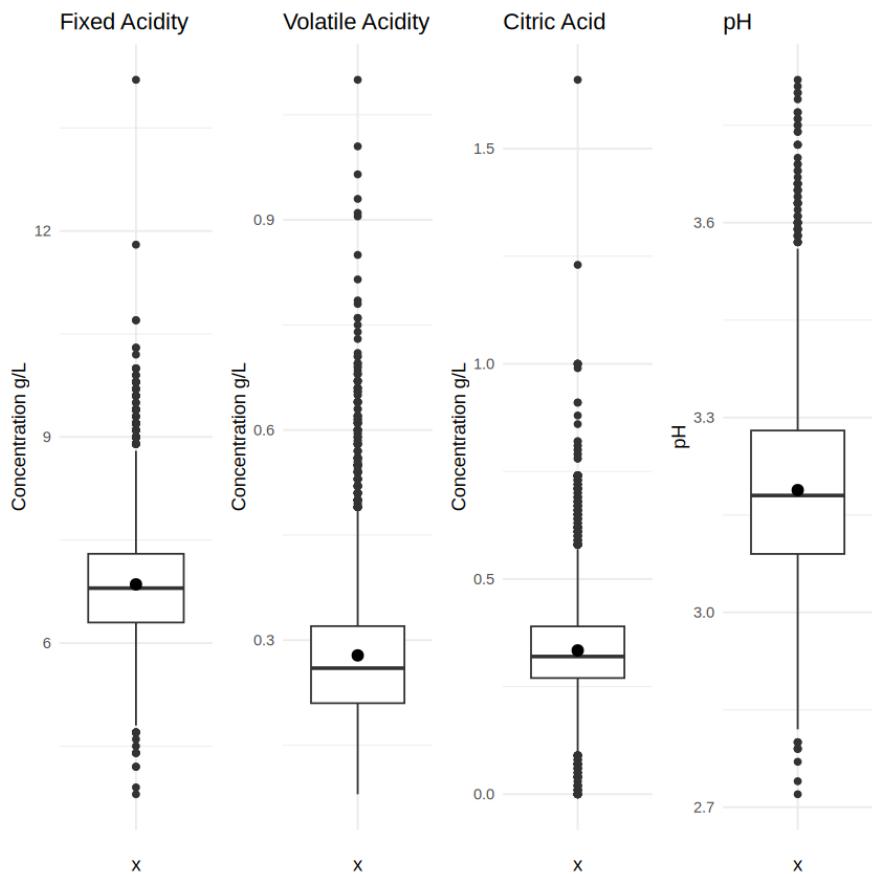
Khảo sát tính chua (acidity) trong rượu trắng



Hình 3.2: Histogram tính chua (acidity) trong rượu trắng.

Nhận xét:

- Fixed và volatile acidity có phân phối (tương đối) bị lệch trái.
- Axit citric tạo thành phân bố biến vì một nhóm rượu vang thường như có nồng độ axit citric gần bằng 0.
- Histogram của pH tương đối đối xứng.
- Có một số ít các ngoại lì trong các biến này.

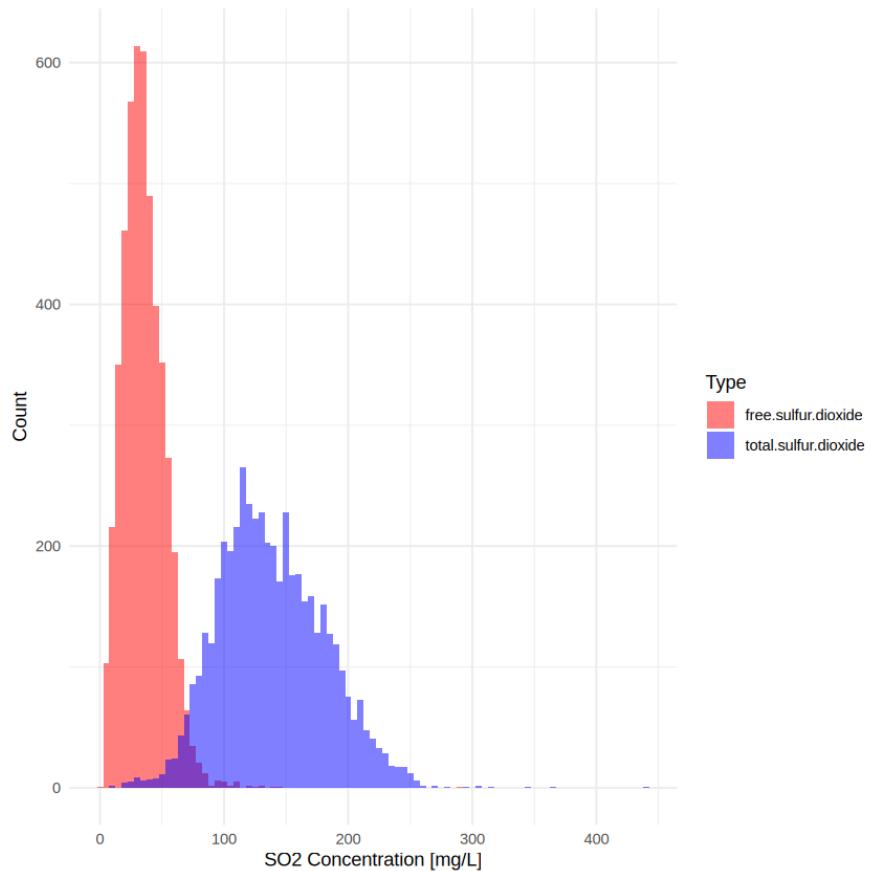


Hình 3.3: Boxplot tính chua (acidity) trong rượu trắng.

Nhận xét:

- Nhìn vào các thông số độ axit trong biểu đồ hộp cho thấy một hình ảnh tương tự.
- Ta có thể thấy đuôi dương dài của nồng độ axit cố định (fixed acide) và dễ bay hơi (volatile acide) và phân phối hẹp hơn đối với axit citric và độ pH.
- Giá trị trung bình của axit citric và pH gần giá trị median hơn là giá trị trung bình của axit cố định (fixed acide) và dễ bay hơi (volatile acide).

Khảo sát hàm lượng lưu huỳnh trong rượu trắng

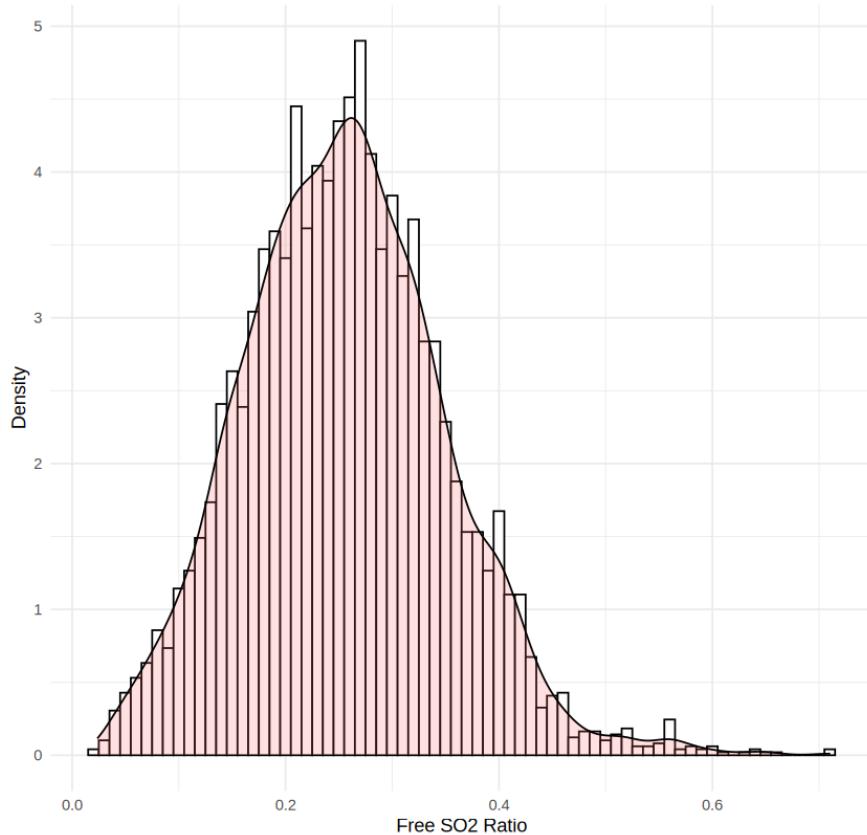


Hình 3.4: Phân phối SO₂ tự do và tổng lượng SO₂ trong rượu.

Nhận xét:

- Nồng độ lưu huỳnh dioxit tự do tập trung hép quanh mức 50 mg/L. Nồng độ lưu huỳnh dioxit tổng thể cho thấy một phân phối tương đối đối xứng.

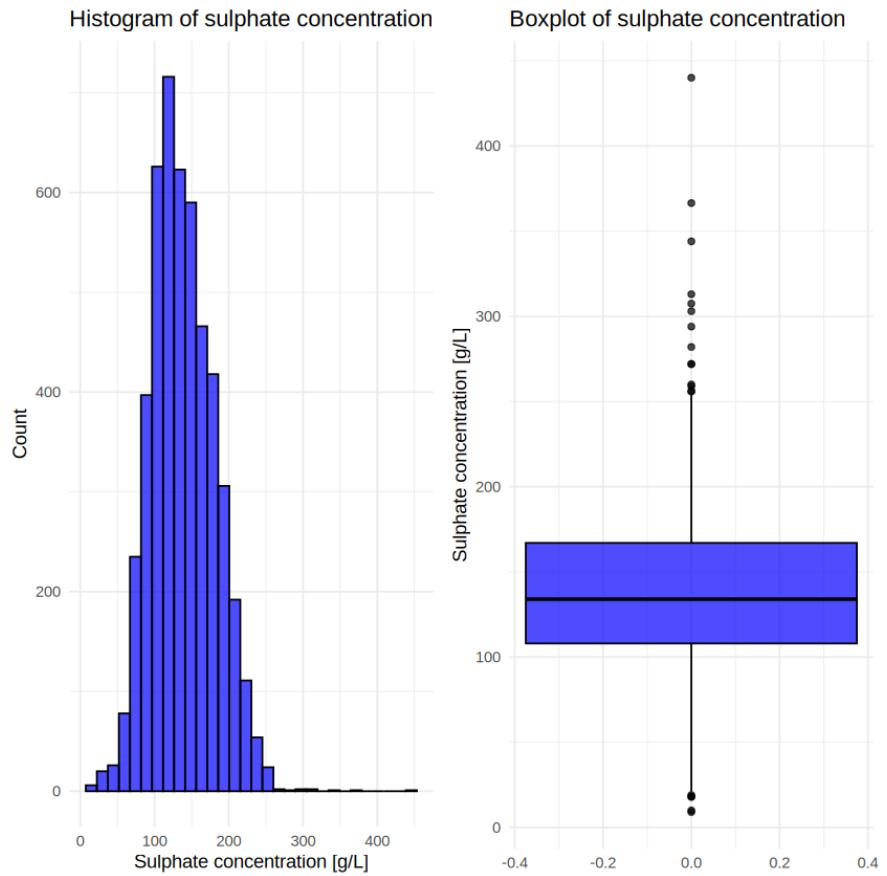
Density Plot of the Ratio of Free to Total Sulfur Dioxide



Hình 3.5: Phân phối tỷ lệ SO₂ tự do và tổng lượng SO₂.

Nhận xét:

- Khi vẽ biểu đồ tỷ lệ giữa lưu huỳnh dioxit tự do và lưu huỳnh dioxit tổng trong rượu vang, người ta có thể thấy rằng khoảng 30% lưu huỳnh dioxit tổng xuất hiện ở dạng tự do. Sự phân bố bị lệch dương với một số loại rượu vang có tỷ lệ cao hơn đáng kể.

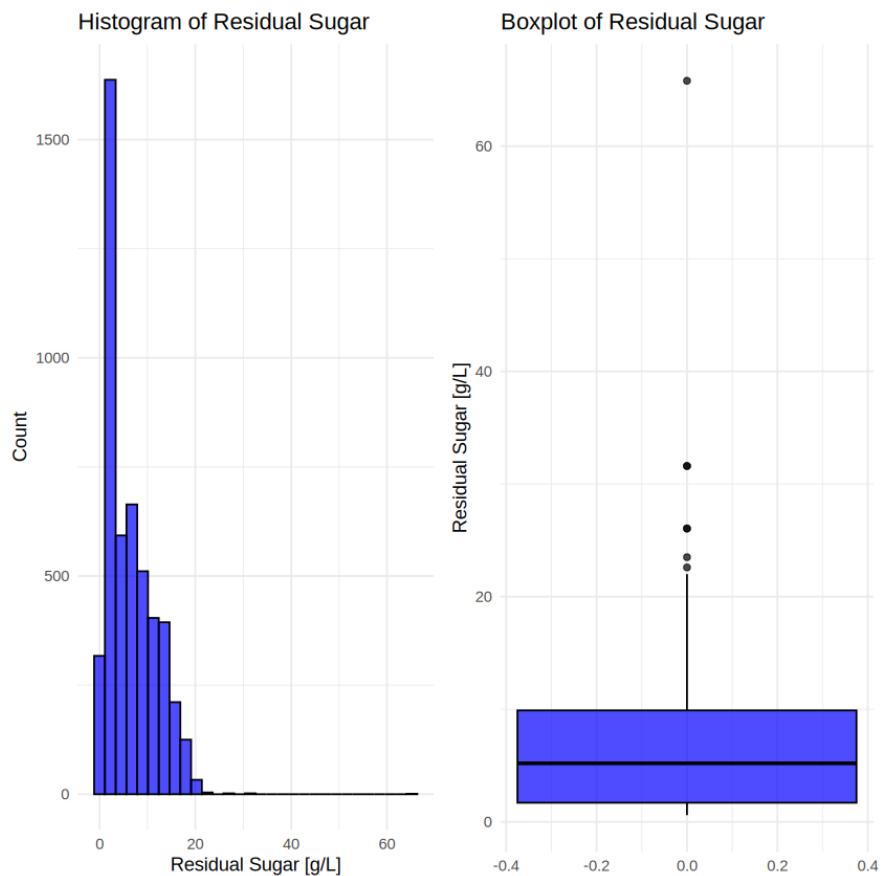


Hình 3.6: Phân phối Lượng muối sunphat trong rượu.

Nhận xét:

- Hầu hết rượu vang trắng có nồng độ sulfat khoảng 0,5 g/L. Có thể thấy ba nhóm ngoại lệ nhỏ trong biểu đồ.

Khảo sát lượng đường còn lại sau khi lên men trong rượu trắng

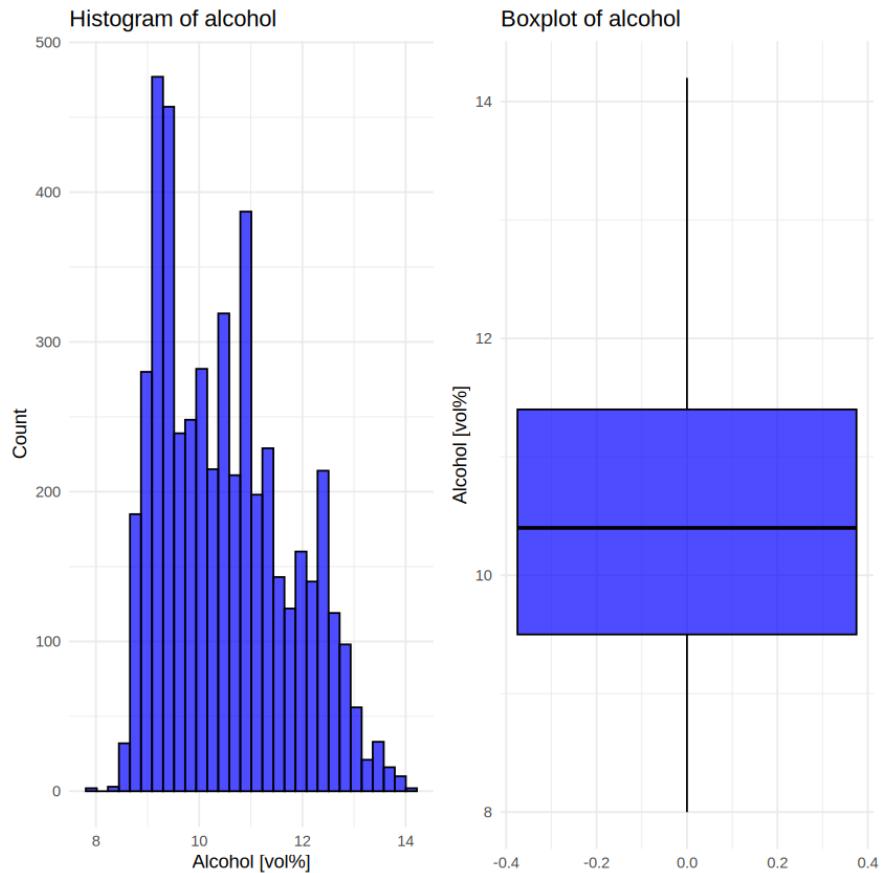


Hình 3.7: Phân phối lượng đường còn lại sau khi lên men trong rượu.

Nhận xét:

- Nhìn chung, rượu vang trắng trong tập dữ liệu có vẻ có nồng độ đường dư thấp gần bằng 0.

Khảo sát phần trăm cồn trong rượu trắng

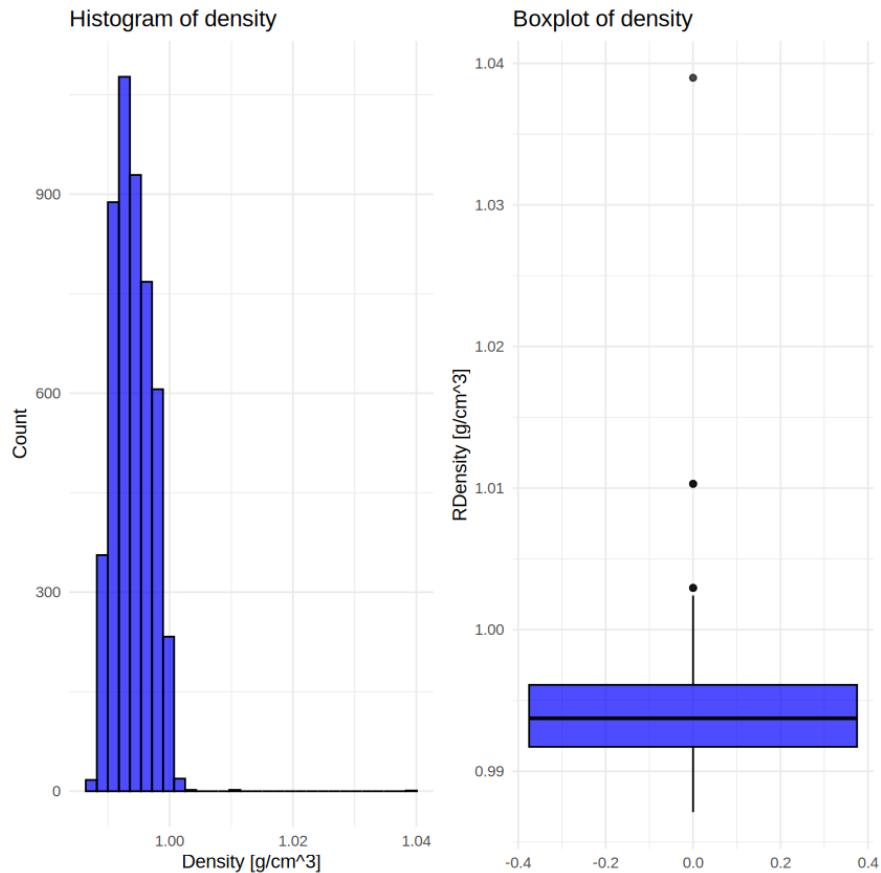


Hình 3.8: Phân phối phần trăm cồn trong rượu.

Nhận xét:

- Hàm lượng cồn của rượu vang trong tập dữ liệu dao động từ 8 đến 15 vol%. Giá trị trung bình nằm trong khoảng 10 vol. Phân phối khá rộng và cho thấy độ lệch dương.

Khảo sát mật độ trong rượu trắng

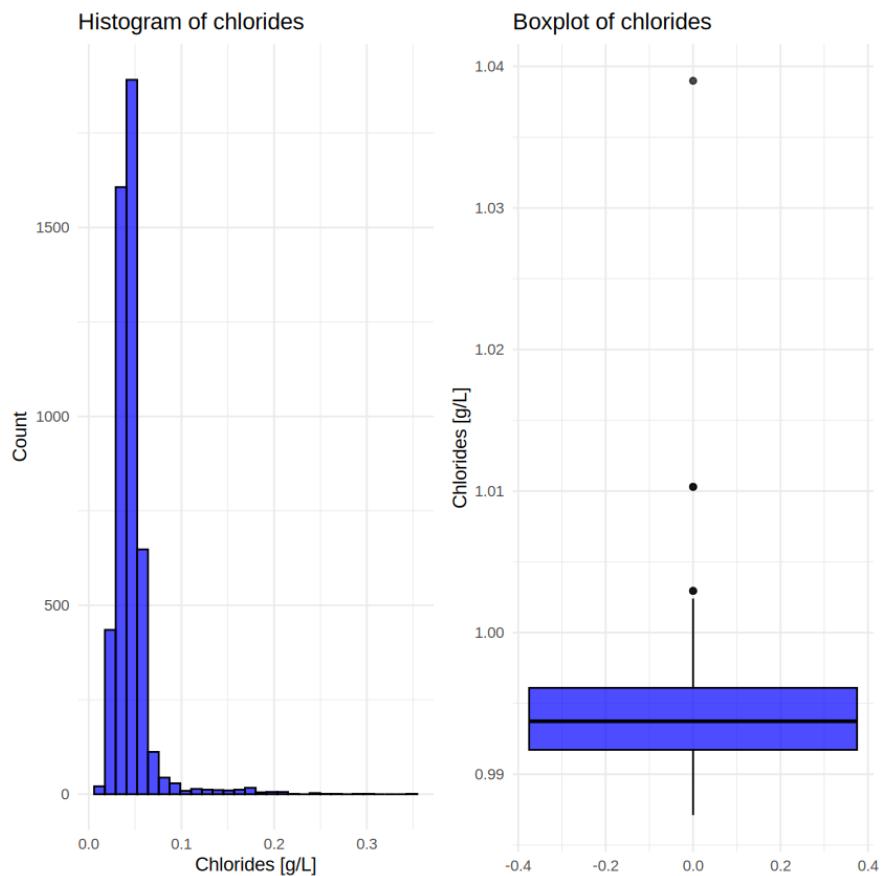


Hình 3.9: Phân phối mật độ rượu.

Nhận xét:

- Tham số mật độ cho thấy sự phân bố rất hẹp với sự thay đổi thấp. Người ta có thể thấy một vài giá trị ngoại lệ trong khoảng 1,01 và 1,04 g/cm³ nhưng hầu hết các loại rượu vang có mật độ trong khoảng 0,99 và 1,00 g/cm³.

Khảo sát lượng muối trong rượu trắng

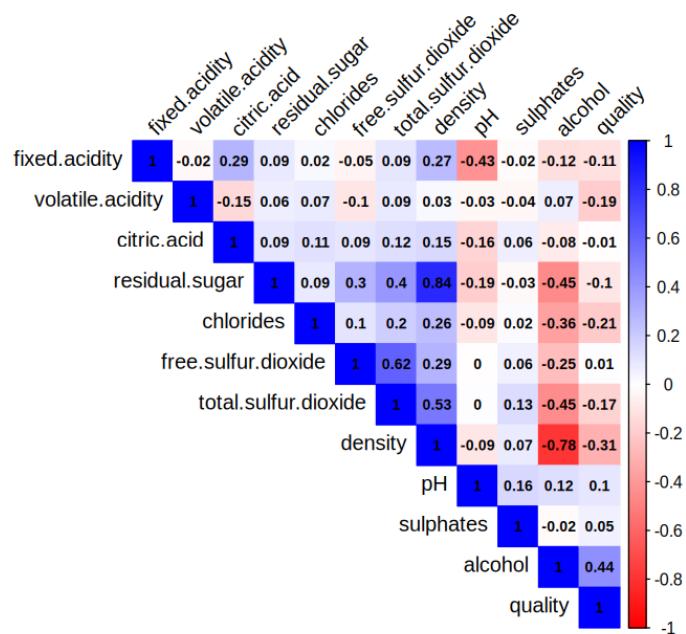


Hình 3.10: Phân phối lượng muối rượu.

Nhận xét:

- Biểu đồ histogram của nồng độ clo cho thấy dữ liệu rượu trắng bị lệch.

Phân tích đa biến



Hình 3.11: Biểu đồ tương quan giữa các biến trong tập dữ liệu rượu trắng.

Chọn ngưỡng là 0.3, ta thấy:

- Nồng độ cồn (alcohol) có ảnh hưởng (thuận) đến chất lượng rượu (chỉ số tương quan 0.436)
- Các biến ‘residual.sugar’ và ‘density’ có tương quan thuận cao 0.83

Chọn ngưỡng là -0.3, ta thấy:

- Mật độ trong rượu (‘density’) có ảnh hưởng (nghịch) đến chất lượng của rượu (chỉ số tương quan -0.307)
- Các biến ‘alcohol’ và ‘density’ có tương quan nghịch cao -0.78

Khảo sát đa công tuyến

Bước 1: Tính toán chỉ số VIF

1	fixed.acidity	volatile.acidity	citric.acid
---	---------------	------------------	-------------

2	2.691435	1.141156	1.165215
3	residual.sugar	chlorides	free.sulfur.dioxide
4	12.644064	1.236822	1.787880
5	total.sulfur.dioxide	density	pH
6	2.239233	28.232546	2.196362
7	sulphates	alcohol	
8	1.138540	7.706957	

Nhận xét:

- Ta có chọn ngưỡng bằng 3

Bước 2: Loại bỏ các biến dựa trên VIF nếu vượt quá ngưỡng

```

1 fixed.acidity      volatile.acidity      citric.acid
2             1.356128                  1.128298          1.159884
3     residual.sugar      chlorides      free.sulfur.dioxide
4             1.435215                  1.203645          1.744627
5 total.sulfur.dioxide      pH      sulphates
6             2.153170                  1.330912          1.056637
7     alcohol
8             1.647117

9
10 Call:
11 lm(formula = quality ~ fixed.acidity + volatile.acidity +
12   citric.acid +
13   residual.sugar + chlorides + free.sulfur.dioxide + total.
14   sulfur.dioxide +
15   pH + sulphates + alcohol, data = wine_quality_white)
16
17 Residuals:
18
19 Coefficients:
20
21              Estimate Std. Error t value Pr(>|t|)    
22 (Intercept) 2.0636371  0.3482321  5.926 3.32e-09 ***
23 fixed.acidity -0.0503197  0.0149092 -3.375 0.000744 ***
24 volatile.acidity -1.9583442  0.1138553 -17.200 < 2e-16 ***
25 citric.acid    -0.0289483  0.0961455 -0.301 0.763360

```

```

25 residual.sugar      0.0256438  0.0025518  10.049 < 2e-16 ***
26 chlorides          -0.9525303  0.5425208  -1.756 0.079194 .
27 free.sulfur.dioxide 0.0047672  0.0008391   5.682 1.41e-08 ***
28 total.sulfur.dioxide -0.0008697  0.0003730  -2.331 0.019771 *
29 pH                  0.1651688  0.0825418   2.001 0.045444 *
30 sulphates          0.4193440  0.0973099   4.309 1.67e-05 ***
31 alcohol             0.3626941  0.0112672  32.190 < 2e-16 ***
32 ---
33 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
34
35 Residual standard error: 0.756 on 4887 degrees of freedom
36 Multiple R-squared:  0.2727 ,    Adjusted R-squared:  0.2713
37 F-statistic: 183.3 on 10 and 4887 DF,  p-value: < 2.2e-16

```

Khảo sát ngoại lai

Ta sử dụng IQR để tìm các điểm ngoại lai và cực ngoại lai:

- Tổng số ngoại lai: 1040
- Tổng số cực ngoại lai: 206

Trong bài toán này, ta sẽ loại bỏ các điểm cực ngoại lai

Chuẩn hóa và phân chia tập dữ liệu

Ta sử dụng box-cox transform và sau đó phân chia tập dữ liệu thành 2 phần: train (80%) và test (20%).

Mô hình hóa hồi quy tuyến tính đa biến

```

1 # Mô hình chẵn dưới
2 model.lb <- lm(quality ~ 1, data = train)
3
4 # Mô hình chẵn trên
5 model.up <- full.lm
6
7 step(full.lm, scope = list(lower = model.lb, upper = model.up),
      direction = "both", trace = FALSE)

```

Kết quả:

```

1 lm(formula = quality ~ fixed.acidity + volatile.acidity +
2   citric.acid +
3   residual.sugar + chlorides + free.sulfur.dioxide + total.
4   sulfur.dioxide +
5   sulphates + alcohol, data = train)
6
7 Coefficients:
8
9   (Intercept)      fixed.acidity      volatile.
10  acidity
11 -1.812e+00       -1.266e-01       -1.267e
12 -04
13  citric.acid      residual.sugar
14  chlorides
15  3.298e-03        4.790e-03       -8.670e
16 -07
17  free.sulfur.dioxide  total.sulfur.dioxide
18  sulphates
19  2.258e-01        2.467e+00       2.396e
20 -04
21  alcohol
22  2.036e+00

```

```

1 wqr_models <- regsubsets(quality ~ volatile.acidity + chlorides
2   + density + pH + sulphates + alcohol, data = train)
3 summary.wqr <- summary(wqr_models)

```

Ta lựa chọn mô hình tốt nhất dựa trên BIC. Kết quả:

```

1 lm(formula = as.formula(formula_str), data = train)
2
3 Residuals:
4
5   Min       1Q     Median       3Q      Max
6   -0.041608 -0.002218  0.000647  0.002862  0.024640
7
8 Coefficients:
9
10             Estimate Std. Error t value Pr(>|t|) 
11 (Intercept) -6.348e-01  4.355e-02 -14.576 < 2e-16 ***
12 fixed.acidity -1.091e-01  2.976e-02  -3.667 0.000249 ***
13 volatile.acidity -1.279e-04  1.118e-05 -11.432 < 2e-16 ***

```

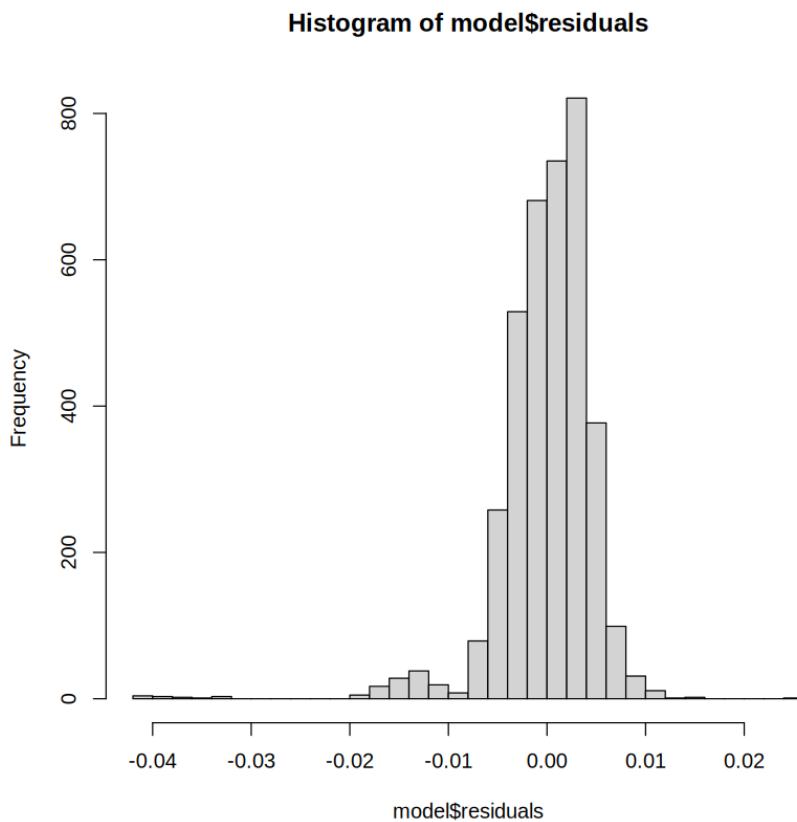
```

12 residual.sugar      4.993e-03  5.319e-04   9.388 < 2e-16 *** 
13 free.sulfur.dioxide 2.482e-01  1.874e-02  13.243 < 2e-16 *** 
14 sulphates          2.405e-04  7.085e-05   3.394 0.000695 *** 
15 alcohol             2.113e+00  7.701e-02  27.431 < 2e-16 *** 
16 --- 
17 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
18 
19 Residual standard error: 0.004705 on 3746 degrees of freedom 
20 Multiple R-squared:  0.2246 ,    Adjusted R-squared:  0.2234 
21 F-statistic: 180.8 on 6 and 3746 DF,  p-value: < 2.2e-16

```

Kiểm định phân phối chuẩn cho các giá trị thặng dư

- H0: Biến thặng dư của mô hình phân phối chuẩn trong một số quần thể.
- H1: Biến thặng dư của mô hình không phân phối chuẩn trong một số quần thể.



Hình 3.12: Histogram của biến thặng dư mô hình.

Kết quả:

```

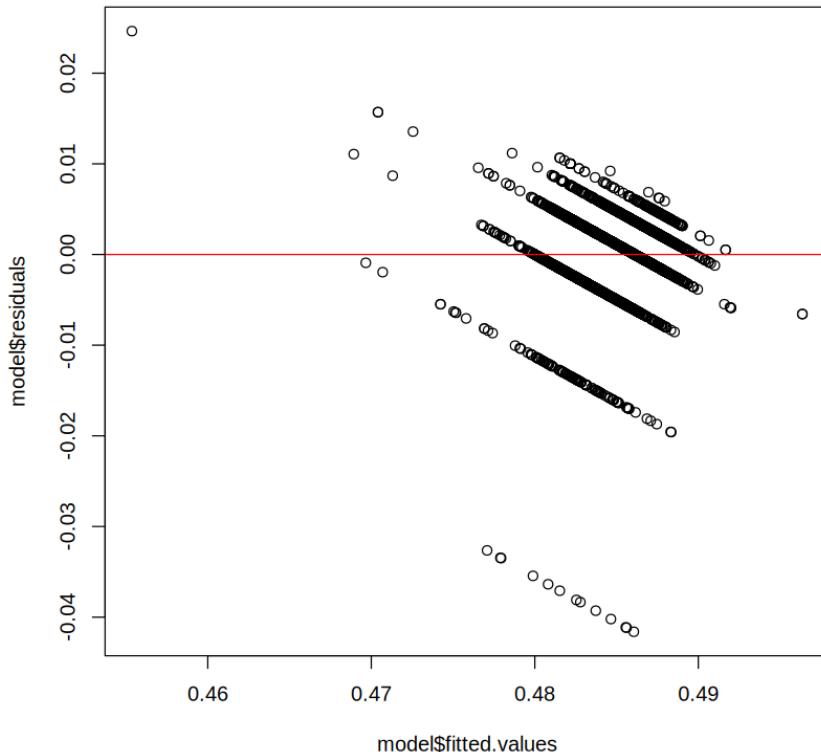
1 Shapiro-Wilk normality test
2
3 data: model$residuals
4 W = 0.8476, p-value < 2.2e-16
5
6 [1] "H0 rejected: the residuals are NOT distributed normally"

```

Như vậy, biến thặng dư không có phân phối chuẩn. Như vậy, các phân tích về sau có thể chưa đủ độ tin cậy. Cần có những biến đổi để cải thiện kết quả phân tích.

Kiểm định phương sai đồng nhất bằng việc sử dụng Biểu đồ scale-location kiểm định giả định hồi quy về phương sai bằng nhau (homoscedasticity), tức là giá trị thặng dư có phương sai bằng với đường hồi quy.

- H0: Các giá trị thặng dư là homoscedastic
- H1: Các giá trị thặng dư là heteroscedastic



Hình 3.13: Biểu đồ Heteroscedasticity.

Kết quả:

```
1 studentized Breusch-Pagan test  
2  
3 data: model  
4 BP = 140.19, df = 6, p-value < 2.2e-16  
5  
6 [1] "H0 rejected: Error variance spreads INCONSTANTLY/  
     generating patterns (Heteroscedasticity)"(Heteroscedasticity  
 ) "
```

Như vậy, ta thấy p-value nhỏ hơn mức ý nghĩa 0.05, ta đủ điều kiện bác bỏ H0. Vậy các giá trị thăng dư là heteroscedastic.

Kết quả dự đoán

Dựa trên quá trình mô hình hóa, ta thu được mô hình

```
1 quality = fixed.acidity + volatile.acidity + residual.sugar +  
    free.sulfur.dioxide + sulphate + alcohol
```

với các hệ số:

```
1 Coefficients:  
2 (Intercept)      fixed.acidity      volatile.acidity  
3           -0.6041273          -0.1045400          -0.0001204  
4 residual.sugar   free.sulfur.dioxide      sulphates  
5           0.0039231          0.2933424          0.0001746  
6         alcohol  
7           2.0023690
```

Điều này có nghĩa là:

- Chất lượng rượu phụ thuộc vào nồng độ cồn, nồng độ cồn càng cao, chất lượng rượu càng tăng.
- Các chỉ số về tính chua khiến chất lượng của rượu bị giảm.
- Lượng đường, muối nhỏ có thể giúp rượu trắng ngon hơn.
- Khí SO2 có tác động tích cực đến chất lượng rượu trắng

3.1.4. Phân tích chất lượng rượu đỏ

Các thông tin thống kê mô tả về bộ dữ liệu

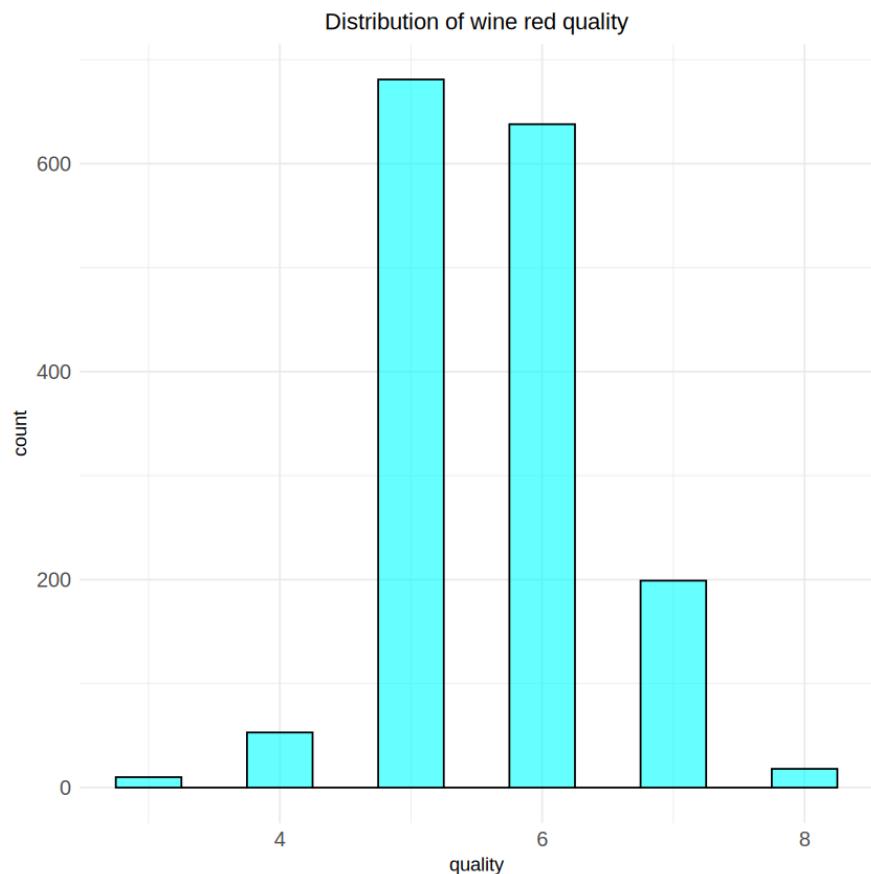
	variable	missing	min	lower	median	upper
	max	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	<chr>					
2						
3	1 fixed.acidity	0	4.6	7.1	7.9	9.2
4	2 volatile.acidity	0	0.12	0.4	0.5	0.6
5	3 citric.acid	0	0	0.1	0.3	0.4
6	4 residual.sugar	0	0.9	1.9	2.2	2.6
7	5 chlorides	0	0.012	0.1	0.1	0.1
		0.611				
8	6 free.sulfur.dioxide	0	1	7	14	21
9	7 total.sulfur.dioxide	0	6	22	38	62
10	8 density	0	0.990	1	1	1
11	9 pH	0	2.74	3.2	3.3	3.4
12	10 sulphates	0	0.33	0.6	0.6	0.7
13	11 alcohol	0	8.4	9.5	10.2	11.1
14	12 quality	0	3	5	6	8

Nhận xét:

-

Phân tích đơn biến

Chất lượng rượu đỏ

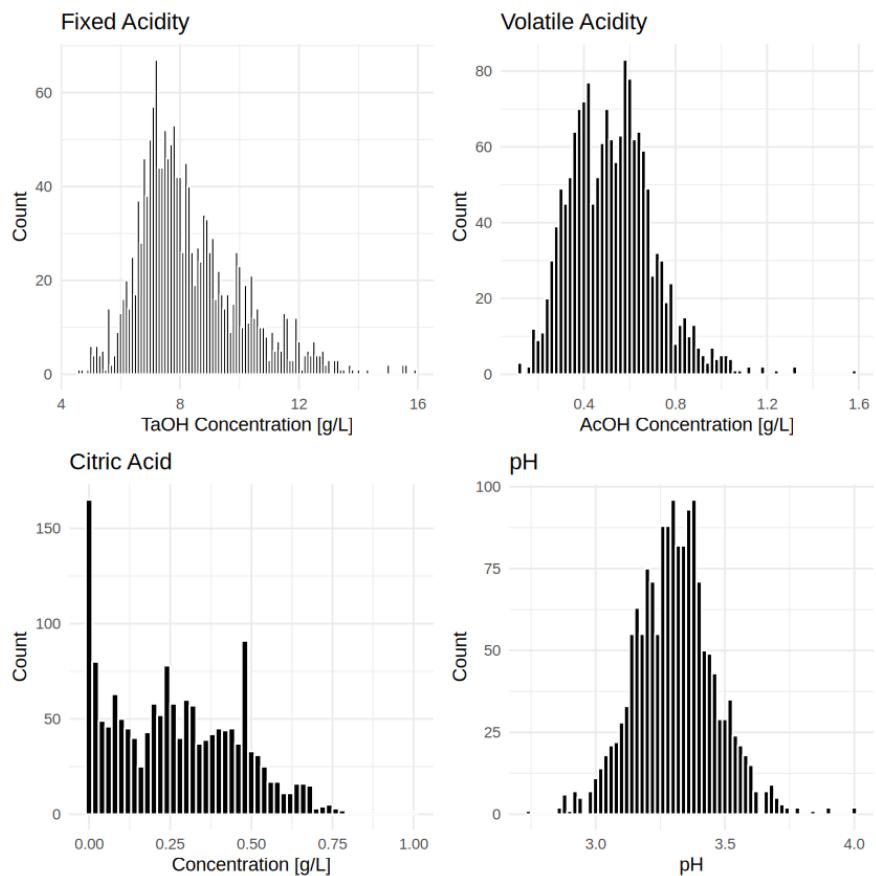


Hình 3.14: Chất lượng rượu đỏ.

Nhận xét:

- Chất lượng rượu có phân phối đối xứng
- Hầu hết chất lượng rượu đỏ nằm ở mức 5, 6
- Không có rượu đỏ nào đạt điểm tuyệt đối
- Chất lượng rượu đỏ tệ nhất có điểm số là 3

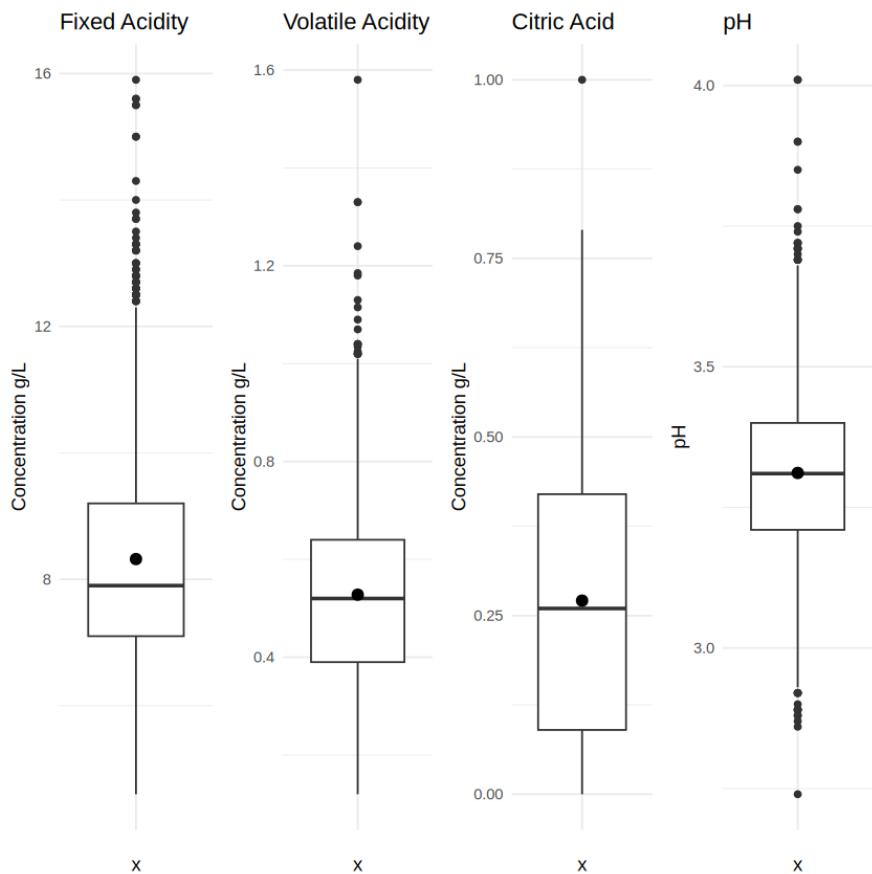
Khảo sát tính chua (acidity) trong rượu đỏ



Hình 3.15: Histogram tính chua (acidity) trong rượu đỏ.

Nhận xét:

- Fixed và volatile acidity có phân phối (tương đối) bị lệch trái.
- Axit citric tạo thành phân bố biến vì một nhóm rượu vang đường như có nồng độ axit citric gần bằng 0.
- Histogram của pH tương đối đối xứng.
- Có một số ít các ngoại lai trong các biến này.

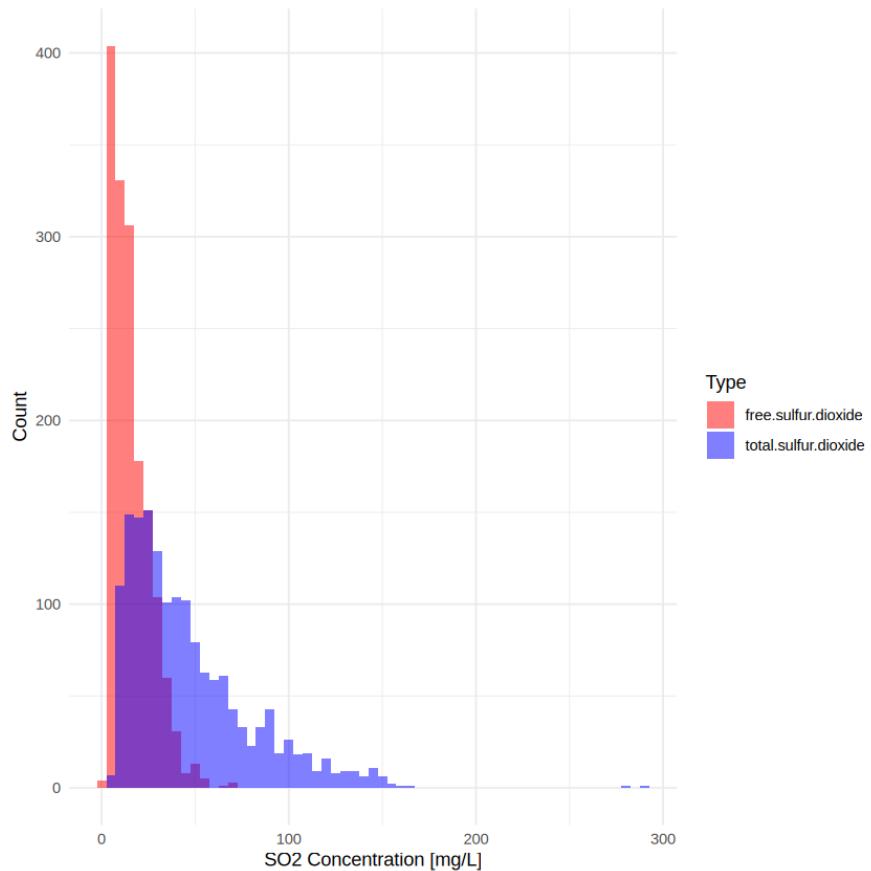


Hình 3.16: Boxplot tính chua (acidity) trong rượu đỏ.

Nhận xét:

- Nhìn vào các thông số độ axit trong biểu đồ hộp cho thấy một hình ảnh tương tự.
- Ta có thể thấy đuôi dương dài của nồng độ axit cố định (fixed acide) và dễ bay hơi (volatile acide) và phân phối hẹp hơn đối với axit citric và độ pH.
- Giá trị trung bình của axit citric và pH gần giá trị median hơn là giá trị trung bình của axit cố định (fixed acide) và dễ bay hơi (volatile acide).

Khảo sát hàm lượng lưu huỳnh trong rượu đỏ

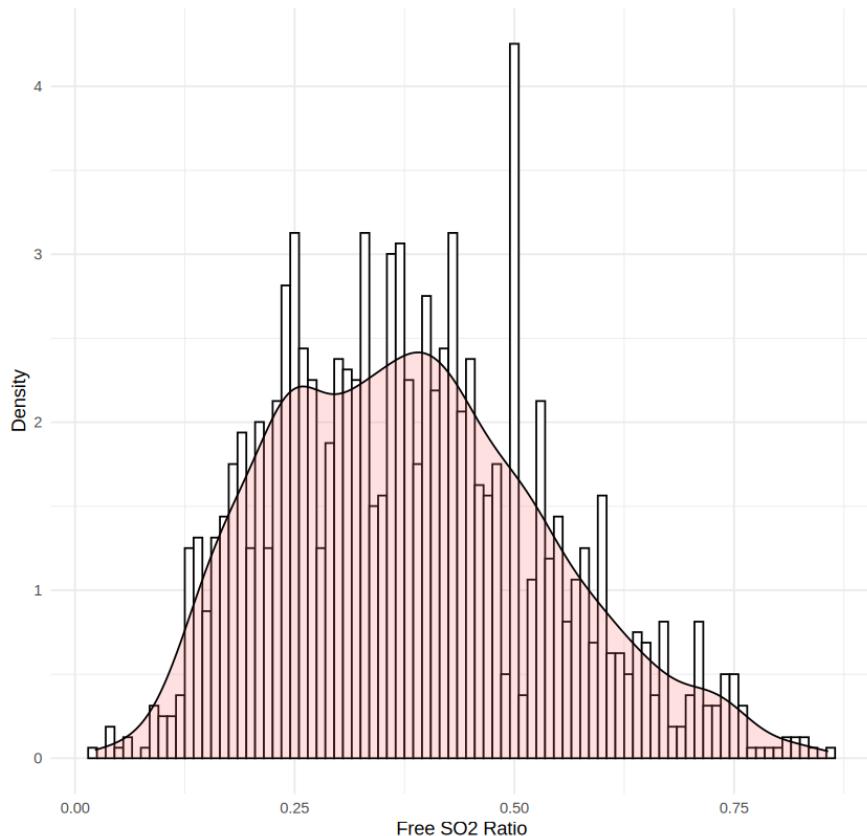


Hình 3.17: Phân phối SO₂ tự do và tổng lượng SO₂ trong rượu.

Nhận xét:

- Nồng độ lưu huỳnh dioxit tự do tập trung hép quanh mức 30 mg/L. Nồng độ lưu huỳnh dioxit tổng thể cho thấy một phân phối bị lệch trái.

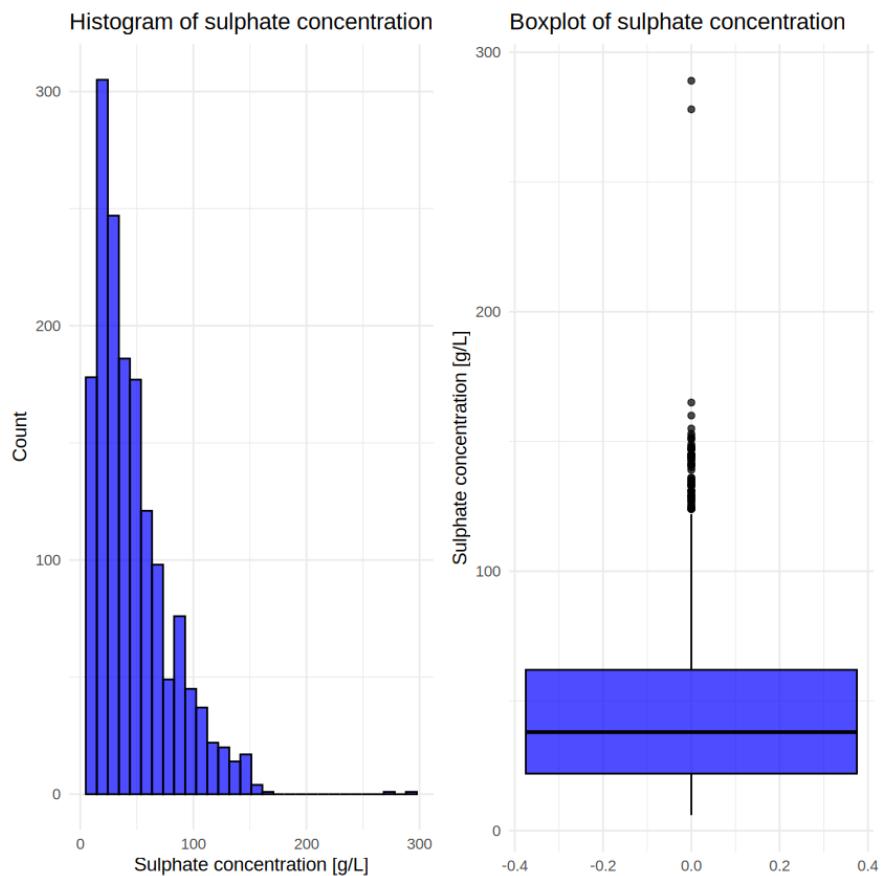
Density Plot of the Ratio of Free to Total Sulfur Dioxide



Hình 3.18: Phân phối tỷ lệ SO₂ tự do và tổng lượng SO₂.

Nhận xét:

- Khi vẽ biểu đồ tỷ lệ giữa lưu huỳnh dioxit tự do và lưu huỳnh dioxit tổng trong rượu vang, người ta có thể thấy rằng khoảng 30% lưu huỳnh dioxit tổng xuất hiện ở dạng tự do. Sự phân bố bị lệch dương với một số loại rượu vang có tỷ lệ cao hơn đáng kể. Điểm đáng chú ý nữa là đỉnh xuất hiện chính xác ở mức 0,5.

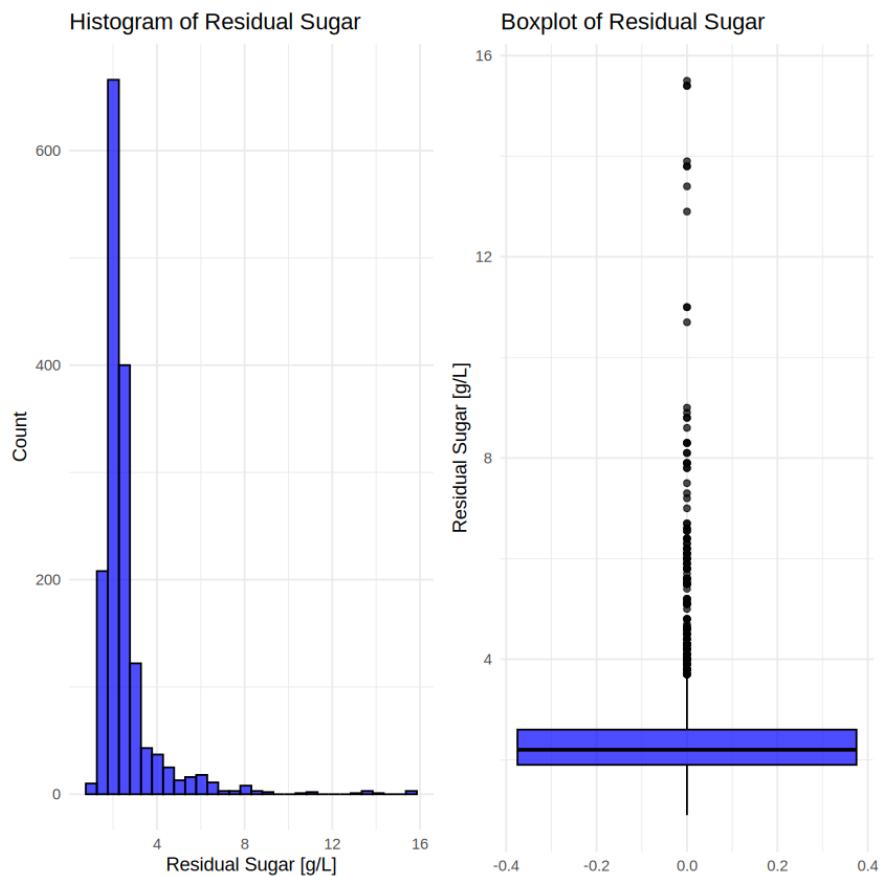


Hình 3.19: Phân phối Lượng muối sunphat trong rượu.

Nhận xét:

- Hầu hết rượu vang đỏ có nồng độ sulfat khoảng 0,5 g/L. Có thể thấy ba nhóm ngoại lệ nhỏ trong biểu đồ.

Khảo sát lượng đường còn lại sau khi lên men trong rượu đỏ

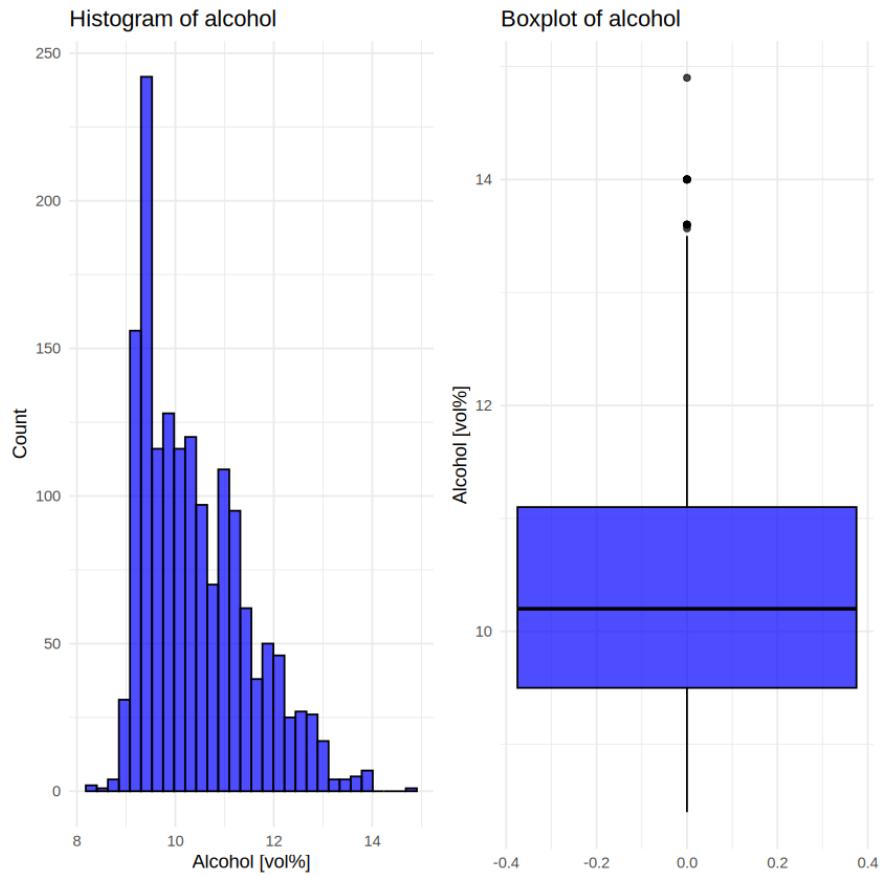


Hình 3.20: Phân phối lượng đường còn lại sau khi lên men trong rượu.

Nhận xét:

- Nhìn chung, rượu vang đỏ trong tập dữ liệu có vẻ có nồng độ đường dư thấp gần bằng 0.

Khảo sát phần trăm cồn trong rượu đỏ

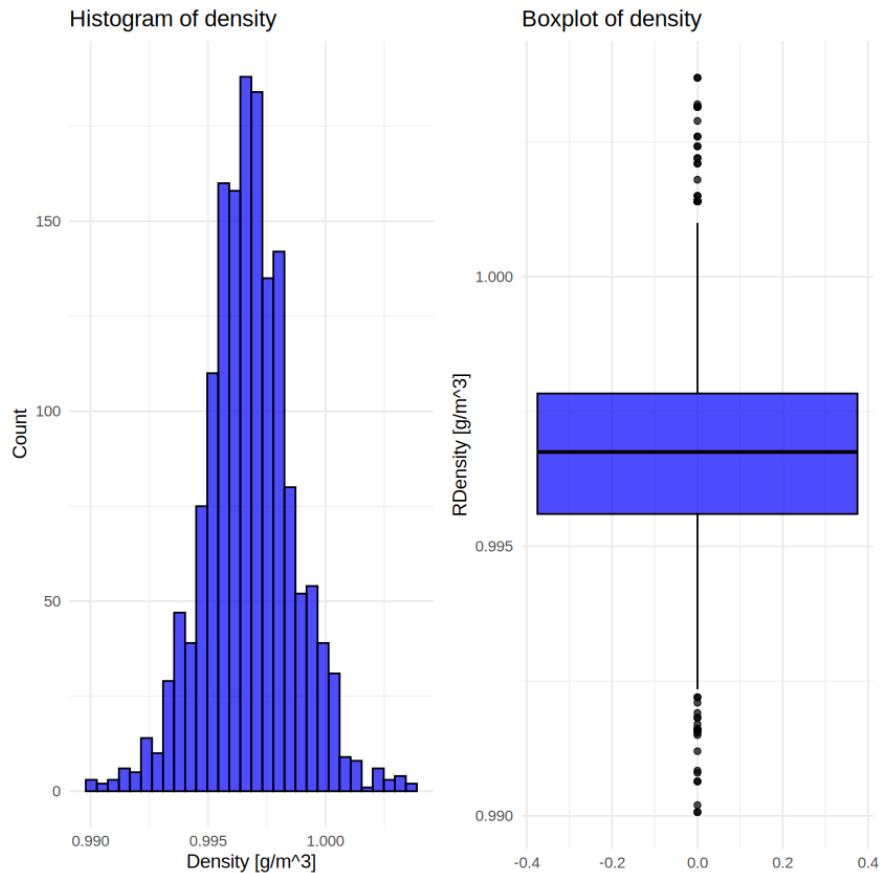


Hình 3.21: Phân phối phần trăm cồn trong rượu.

Nhận xét:

- Hàm lượng cồn của rượu vang đỏ trong tập dữ liệu dao động từ 8 đến 15 vol%. Giá trị trung bình nằm trong khoảng 10 vol. Phân phối khá rộng và cho thấy độ lệch dương.

Khảo sát mật độ trong rượu đỏ

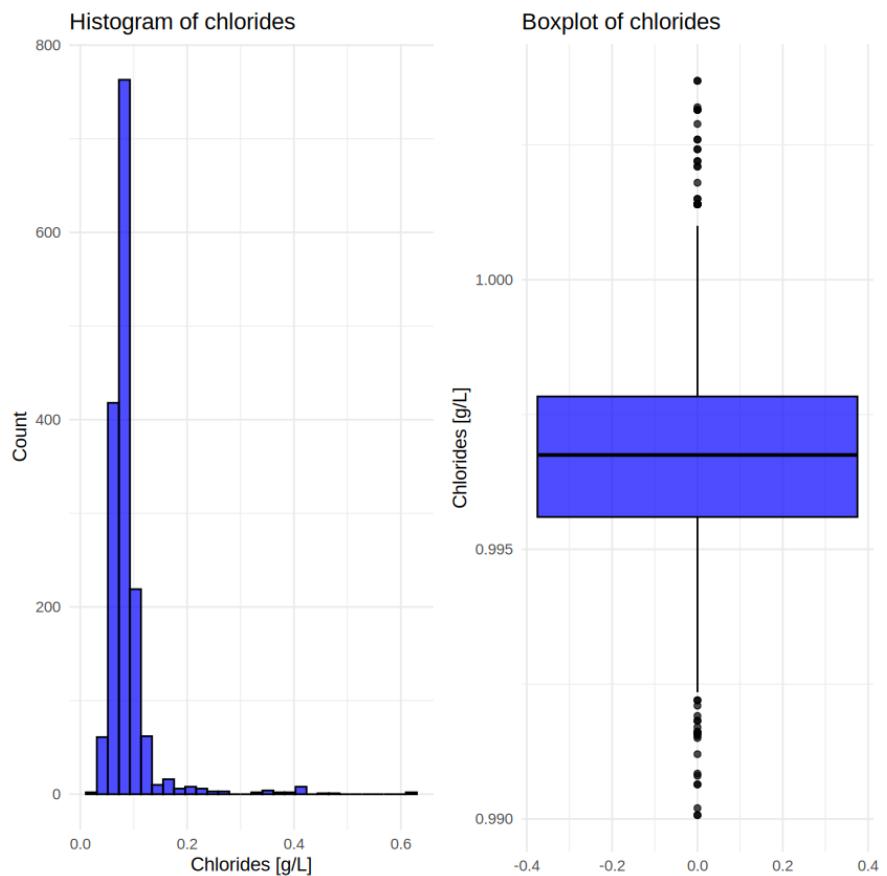


Hình 3.22: Phân phối mật độ rượu.

Nhận xét:

- Tham số mật độ cho thấy sự phân bố rất hẹp với sự thay đổi thấp. Người ta có thể thấy một vài giá trị ngoại lệ trong khoảng 1,01 và 1,04 g/cm³ nhưng hầu hết các loại rượu vang có mật độ trong khoảng 0,99 và 1,00 g/cm³.

Khảo sát lượng muối trong rượu đỏ

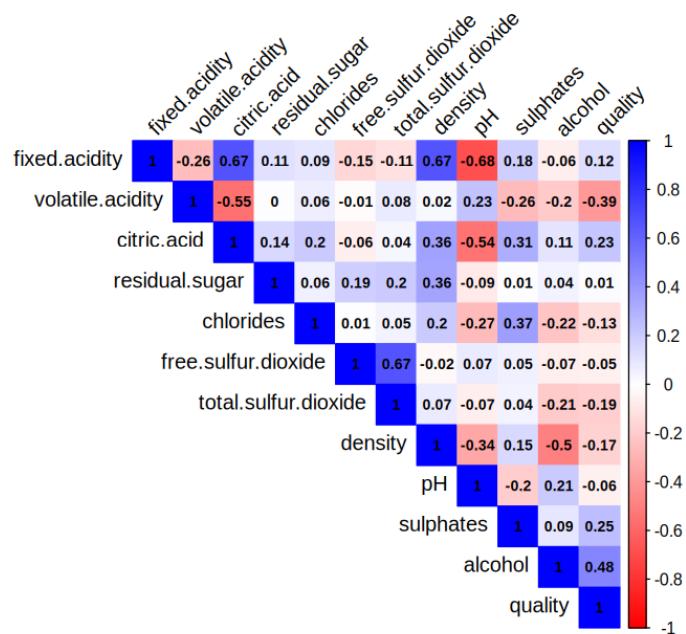


Hình 3.23: Phân phối lượng muối rượu.

Nhận xét:

- Biểu đồ histogram của nồng độ clo cho thấy dữ liệu rượu đỏ tương đối cân bằng.

Phân tích đa biến



Hình 3.24: Biểu đồ tương quan giữa các biến trong tập dữ liệu rượu đỏ.

Chọn ngưỡng là 0.3, ta thấy:

- Nồng độ cồn (alcohol) có ảnh hưởng (thuận) đến chất lượng rượu (chỉ số tương quan 0.476)
- Các biến ‘residual.sugar’ và ‘density’ có tương quan thuận thấp 0.35
- Biến fixed.acidity và citric.acid có tương quan dương mạnh, 0.671

Chọn ngưỡng là -0.3, ta thấy:

- Mật độ trong rượu (‘density’) có ảnh hưởng (nghịch) đến pH của rượu (chỉ số tương quan -0.34) và alcohol (-0.496)

Khảo sát đa công tuyến

Bước 1: Tính toán chỉ số VIF

1	fixed.acidity	volatile.acidity	citric.acid
---	---------------	------------------	-------------

2	7.767512	1.789390	3.128022
3	residual.sugar	chlorides free.sulfur.dioxide	
4	1.702588	1.481932	1.963019
5	total.sulfur.dioxide	density	pH
6	2.186813	6.343760	3.329732
7	sulphates	alcohol	
8	1.429434	3.031160	

Nhận xét:

- Ta có chọn ngưỡng bằng 3

Bước 2: Loại bỏ các biến dựa trên VIF nếu vượt quá ngưỡng

```

1 volatile.acidity          citric.acid      residual.sugar
2           1.784963          2.780557      1.386375
3      chlorides  free.sulfur.dioxide total.sulfur.dioxide
4           1.401232          1.939209      2.069396
5      density                  pH      sulphates
6           2.430096          1.610775      1.396382
7      alcohol
8           2.136067

9
10 Call:
11 lm(formula = quality ~ volatile.acidity + citric.acid +
12   residual.sugar +
13   chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
14   density + pH + sulphates + alcohol, data = wine_quality_red
15 )
16
17 Residuals:
18
19 Coefficients:
20
21             Estimate Std. Error t value Pr(>|t|)
22 (Intercept) 6.1795700 13.4367180 0.460 0.6456
23 volatile.acidity -1.0777894 0.1209486 -8.911 < 2e-16 ***
24 citric.acid -0.1353226 0.1387582 -0.975 0.3296
25 residual.sugar 0.0101047 0.0135372 0.746 0.4555

```

```

25 chlorides           -1.9684566  0.4076978  -4.828  1.51e-06 ***
26 free.sulfur.dioxide 0.0045916  0.0021580   2.128   0.0335 *
27 total.sulfur.dioxide -0.0034272  0.0007089  -4.835  1.46e-06 ***
28 density              -1.5167406  13.3889717  -0.113   0.9098
29 pH                   -0.5462340  0.1332577  -4.099  4.36e-05 ***
30 sulphates            0.8995900  0.1130053   7.961  3.23e-15 ***
31 alcohol               0.2900579  0.0222316  13.047 < 2e-16 ***
32 ---
33 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
34
35 Residual standard error: 0.648 on 1588 degrees of freedom
36 Multiple R-squared:  0.3602,    Adjusted R-squared:  0.3561
37 F-statistic: 89.39 on 10 and 1588 DF,  p-value: < 2.2e-16

```

Khảo sát ngoại lai

Ta sử dụng IQR để tìm các điểm ngoại lai và cực ngoại lai:

- Tổng số ngoại lai: 393
- Tổng số cực ngoại lai: 160

Trong bài toán này, ta sẽ loại bỏ các điểm cực ngoại lai

Chuẩn hóa và phân chia tập dữ liệu

Ta sử dụng box-cox transform và sau đó phân chia tập dữ liệu thành 2 phần: train (80%) và test (20%).

Mô hình hóa hồi quy tuyến tính đa biến

```

1 # Mô hình chặn dưới
2 model.lb <- lm(quality ~ 1, data = train)
3
4 # Mô hình chặn trên
5 model.up <- full.lm
6
7 step(full.lm, scope = list(lower = model.lb, upper = model.up),
      direction = "both", trace = FALSE)

```

Kết quả:

```

1 lm(formula = quality ~ volatile.acidity + chlorides + density +
2   pH + sulphates + alcohol, data = train)
3
4 Coefficients:
5   (Intercept)  volatile.acidity          chlorides
6   density
7   -2.489e-01      -2.356e-04      -2.089e-06
8   -2.798e-01
9   pH           sulphates            alcohol
10  -1.626e-01      2.639e-03      1.629e+00

```

```

1 wqr_models <- regsubsets(quality ~ volatile.acidity + chlorides +
2   + density + pH + sulphates + alcohol, data = train)
3 summary.wqr <- summary(wqr_models)

```

Ta lựa chọn mô hình tốt nhất dựa trên BIC. Kết quả:

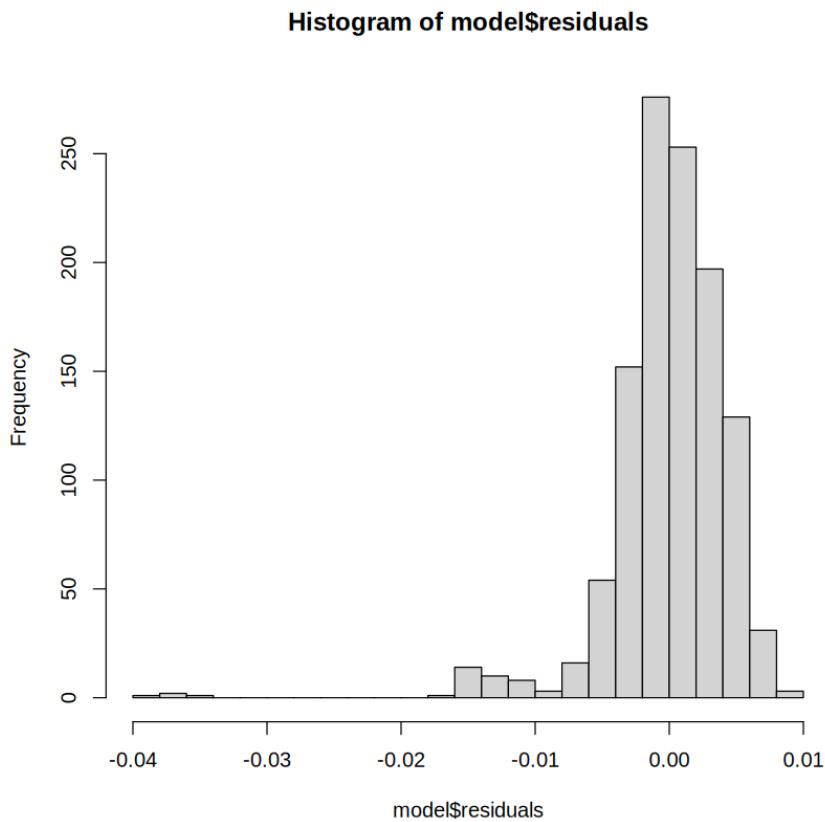
```

1 lm(formula = as.formula(formula_str), data = train)
2
3 Residuals:
4   Min       1Q   Median       3Q      Max
5 -0.039242 -0.001816  0.000235  0.002780  0.009180
6
7 Coefficients:
8   Estimate Std. Error t value Pr(>|t|)
9 (Intercept) -2.579e-01  8.929e-02 -2.888 0.003952 ***
10 volatile.acidity -2.447e-04  4.915e-05 -4.979 7.38e-07 ***
11 density      -3.156e-01  8.656e-02 -3.646 0.000279 ***
12 pH           -1.648e-01  3.290e-02 -5.008 6.36e-07 ***
13 sulphates     2.545e-03  2.811e-04   9.054 < 2e-16 ***
14 alcohol        1.649e+00  1.773e-01   9.300 < 2e-16 ***
15 ---
16 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
17
18 Residual standard error: 0.004389 on 1145 degrees of freedom
19 Multiple R-squared:  0.2724,    Adjusted R-squared:  0.2693
20 F-statistic: 85.75 on 5 and 1145 DF, p-value: < 2.2e-16

```

Kiểm định phân phối chuẩn cho các giá trị thặng dư

- H0: Biến thặng dư của mô hình phân phối chuẩn trong một số quần thể.
- H1: Biến thặng dư của mô hình không phân phối chuẩn trong một số quần thể.



Hình 3.25: Histogram của biến thặng dư mô hình.

Kết quả:

```

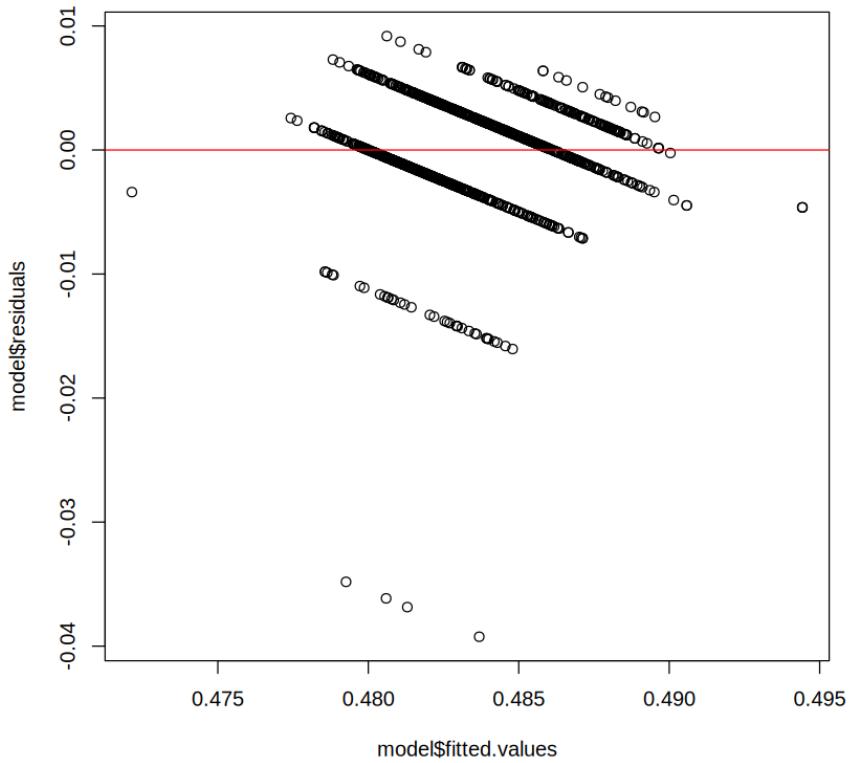
1 Shapiro-Wilk normality test
2
3 data:  model$residuals
4 W = 0.81862, p-value < 2.2e-16
5
6 [1] "H0 rejected: the residuals are NOT distributed normally"

```

Như vậy, biến thặng dư không có phân phối chuẩn. Như vậy, các phân tích về sau có thể chưa đủ độ tin cậy. Cần có những biến đổi để cải thiện kết quả phân tích.

Kiểm định phương sai đồng nhất bằng việc sử dụng Biểu đồ scale-location kiểm định giả định hồi quy về phương sai bằng nhau (homoscedasticity), tức là giá trị thặng dư có phương sai bằng với đường hồi quy.

- H0: Các giá trị thặng dư là homoscedastic
- H1: Các giá trị thặng dư là heteroscedastic



Hình 3.26: Biểu đồ Heteroscedasticity.

Kết quả:

```

1 studentized Breusch-Pagan test
2
3 data: model
4 BP = 15.396, df = 5, p-value = 0.008798
5
6 [1] "H0 rejected: Error variance spreads INCONSTANTLY/
      generating patterns (Heteroscedasticity)"

```

Như vậy, ta thấy p-value nhỏ hơn mức ý nghĩa 0.05, ta đủ điều kiện bác bỏ H0. Vậy các giá trị thặng dư là heteroscedastic

Kết quả dự đoán

Dựa trên quá trình mô hình hóa, ta thu được mô hình

```
1 quality = volatile.acidity + density + pH + sulphates + alcohol
```

với các hệ số:

```
1 Coefficients:
2   (Intercept)  volatile.acidity           density
3                   -0.3675126          -0.0002588       -0.1598464
4                   -0.0902720
5   sulphates      alcohol
6       0.0021522        1.8035125
```

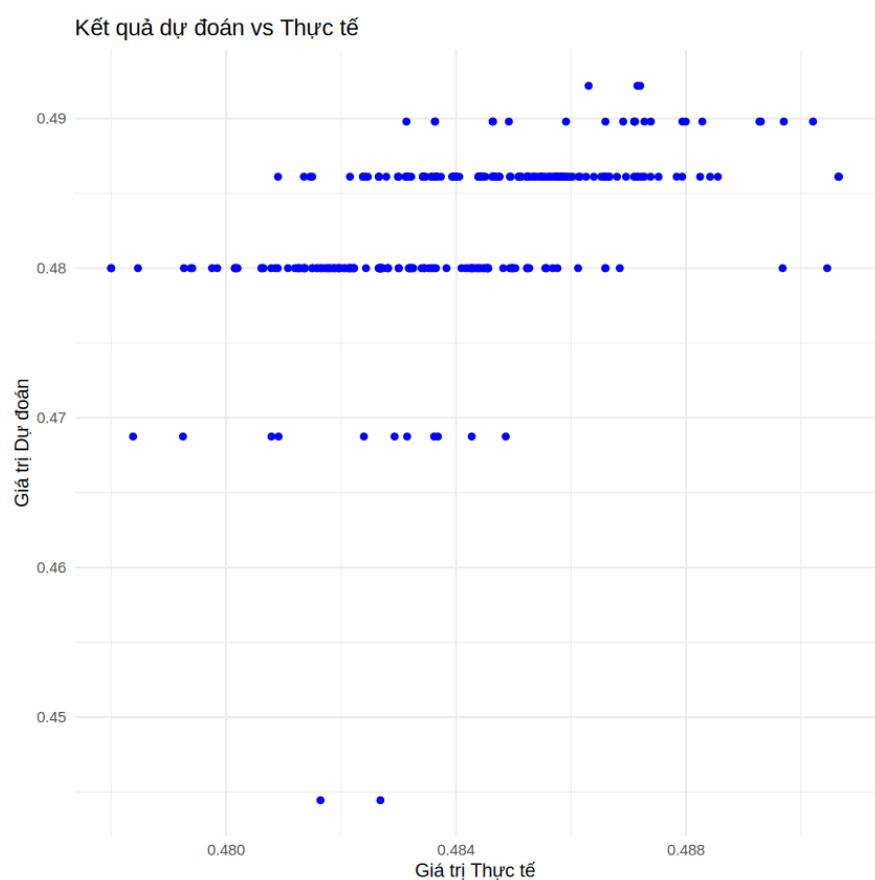
Điều này có nghĩa là:

- Chất lượng rượu phụ thuộc vào nồng độ cồn, nồng độ cồn càng cao, chất lượng rượu càng tăng
- Các chỉ số về tính chua, mật độ rượu, độ pH khiến chất lượng giảm.

Ta sử dụng mô hình để dự đoán kết quả:

- "MSE: 2.6e-05"
- "RMSE: 0.005062"
- "MAE: 0.003118"
- "Correlation: 0.464513"
- " R^2 between y_pred & y_true: 0.215773"

Trực quan hóa:



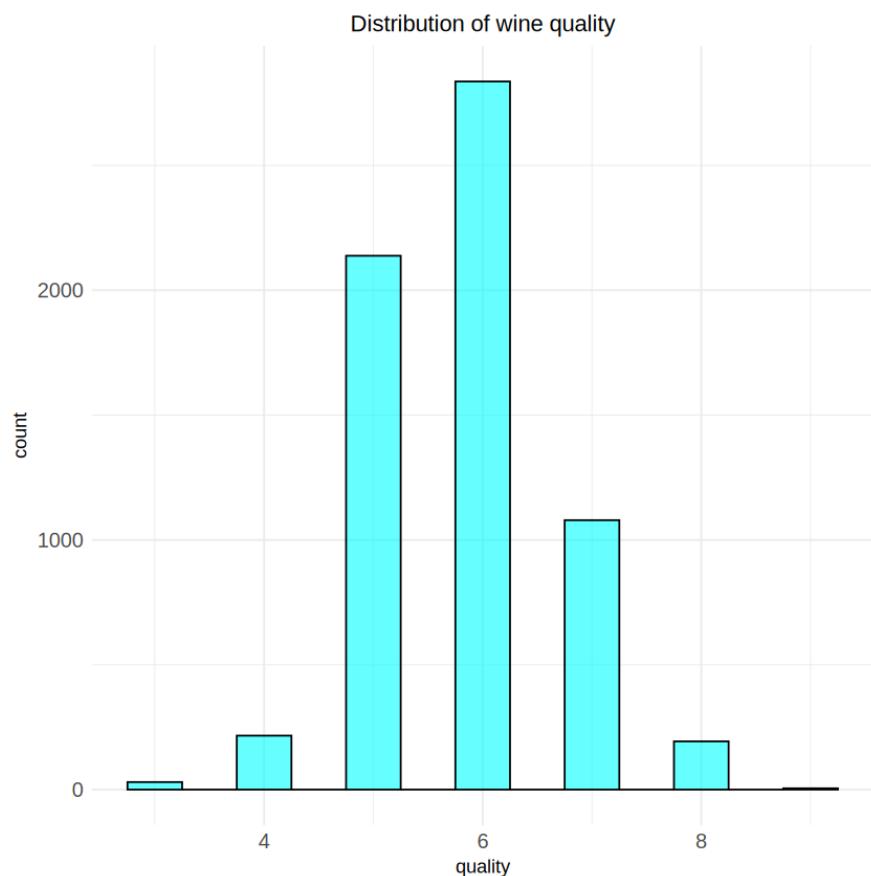
Hình 3.27: Kết quả dự đoán trên bộ dữ liệu chất lượng rượu đỏ.

3.1.5. Phân tích chất lượng rượu (bao gồm nhiều tố màu sắc)

Các thông tin thống kê mô tả về tập dữ liệu

Phân tích đơn biến

Chất lượng rượu

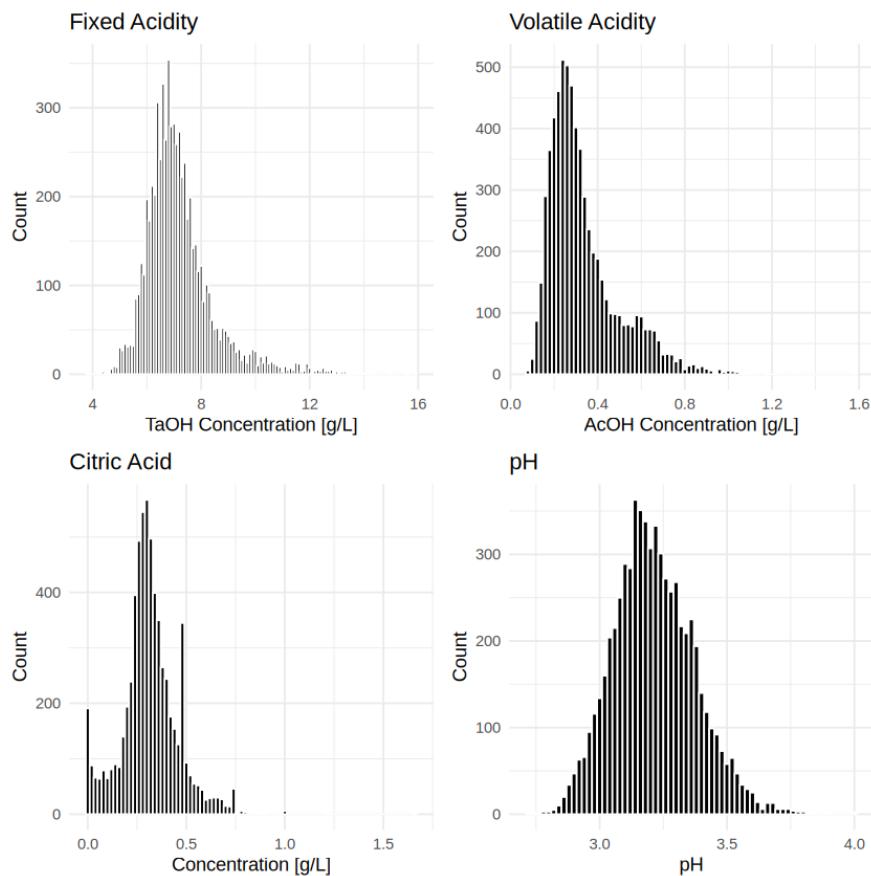


Hình 3.28: Chất lượng rượu.

Nhận xét:

- Chất lượng rượu có phân phối đối xứng
- Hầu hết chất lượng rượu dở nằm ở mức 5, 6
- Không có rượu dở nào đạt điểm tuyệt đối
- Chất lượng rượu dở tệ nhất có điểm số là 3

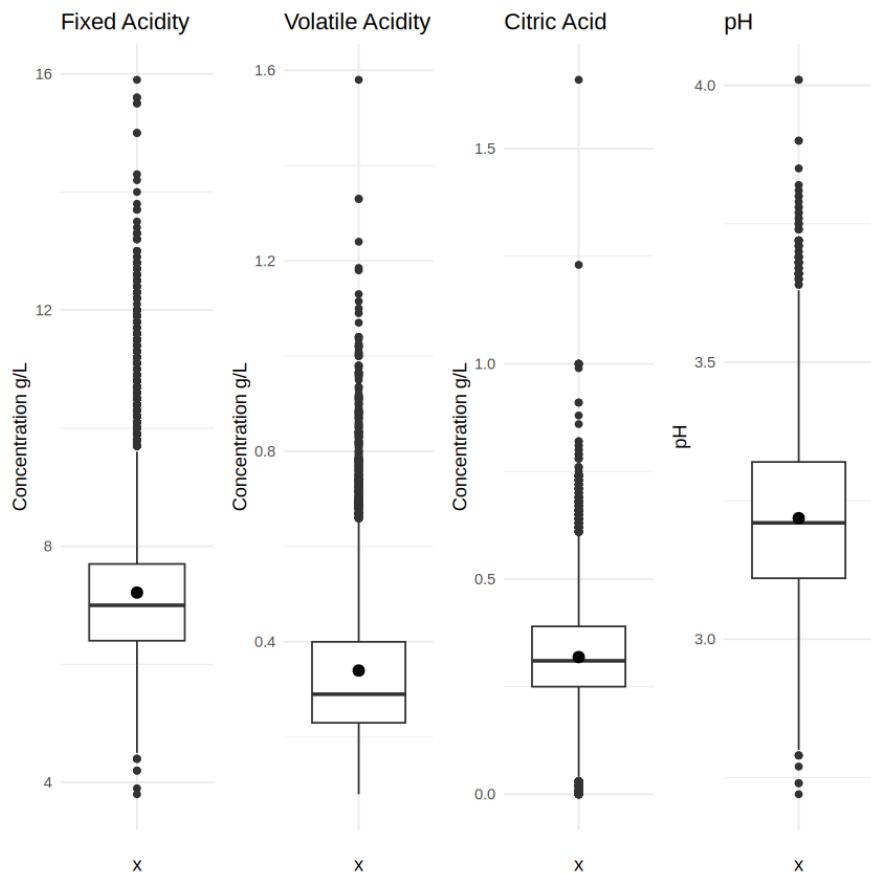
Khảo sát tính chua (acidity) trong rượu



Hình 3.29: Histogram tính chua (acidity) trong rượu.

Nhận xét:

- Độ axit cố định (fixed acidity) và dẽ bay hơi (volatile acidity) cho thấy sự phân bố lệch dương (lệch phải). Axit xitric tạo thành sự phân bố đỉnh cạnh vì một nhóm rượu vang đường như có nồng độ axit xitric gần bằng 0. Biểu đồ pH có vẻ đối xứng hơn. Chỉ có một vài giá trị ngoại lệ có mặt xuất hiện ở các đặc trưng này.

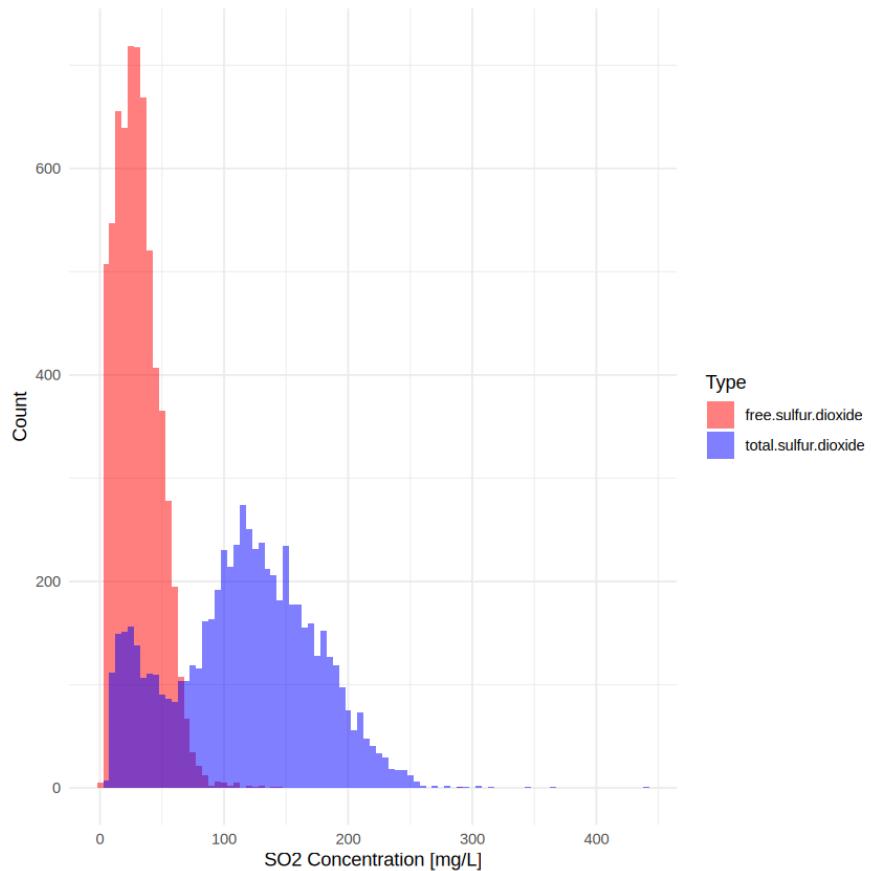


Hình 3.30: Boxplot tính chua (acidity) trong rượu.

Nhận xét:

- Nhìn vào các thông số độ axit trong biểu đồ hộp cho thấy một hình ảnh tương tự. Người ta có thể thấy đuôi dương dài của nồng độ axit cố định và dễ bay hơi và phân phối hẹp hơn đối với axit citric và độ pH. Quan sát này cũng được xác nhận bởi các giá trị trung bình được hiển thị bằng dấu chấm đen trong biểu đồ. Chúng gần với các giá trị trung bình tương ứng đối với axit citric và độ pH hơn là đối với độ axit cố định và dễ bay hơi.

Khảo sát hàm lượng lưu huỳnh trong rượu

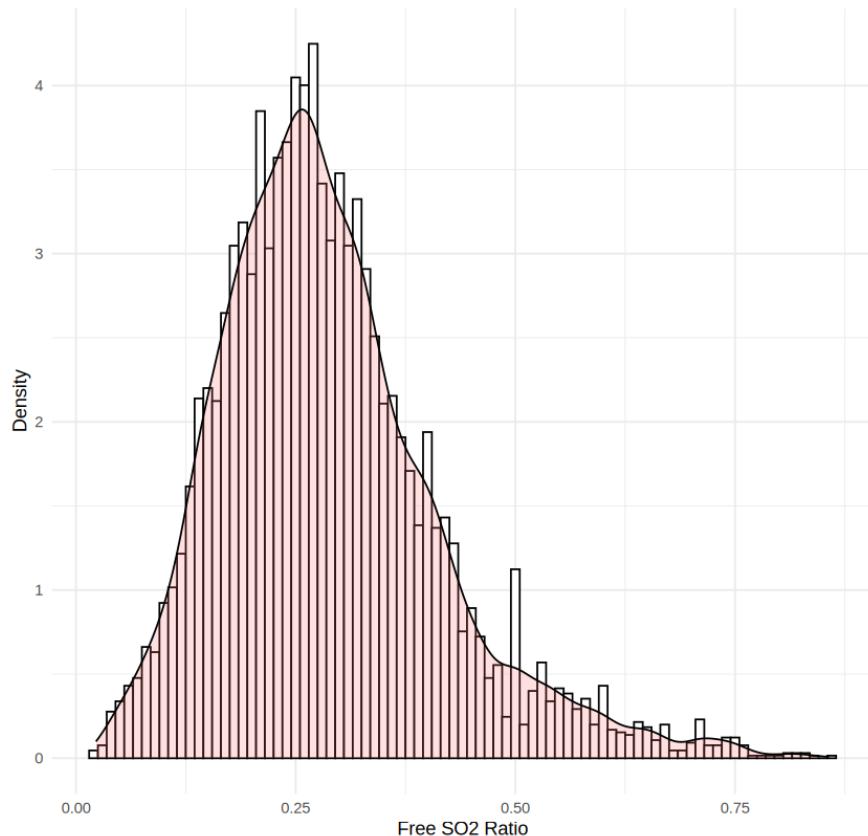


Hình 3.31: Phân phối SO₂ tự do và tổng lượng SO₂ trong rượu.

Nhận xét:

- Nồng độ lưu huỳnh dioxit tự do tập trung hép quanh mức 30 mg/L. Nồng độ lưu huỳnh dioxit tổng thể cho thấy dấu hiệu lưỡng cực với các đỉnh quanh mức 20 và 120 mg/L.

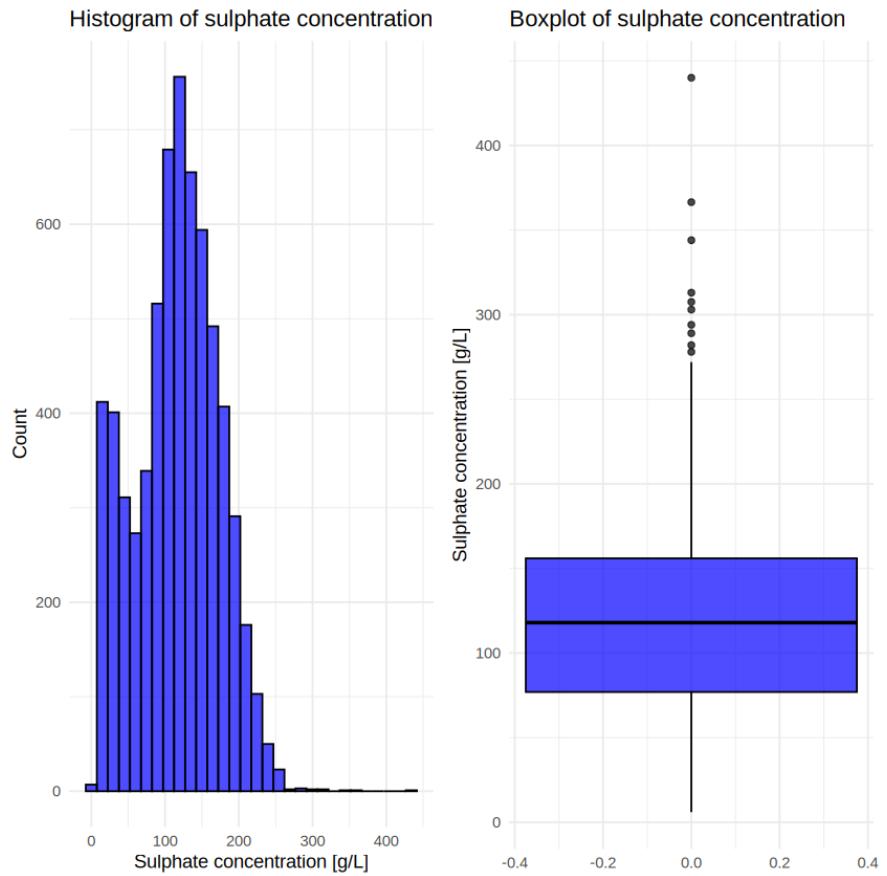
Density Plot of the Ratio of Free to Total Sulfur Dioxide



Hình 3.32: Phân phối tỷ lệ SO₂ tự do và tổng lượng SO₂.

Nhận xét:

- Khi vẽ biểu đồ tỷ lệ giữa lưu huỳnh dioxit tự do và lưu huỳnh dioxit tổng trong rượu vang, người ta có thể thấy rằng khoảng 30% lưu huỳnh dioxit tổng xuất hiện ở dạng tự do. Sự phân bố bị lệch dương với một số loại rượu vang có tỷ lệ cao hơn đáng kể. Điểm đáng chú ý nữa là đỉnh xuất hiện chính xác ở mức 0,5.

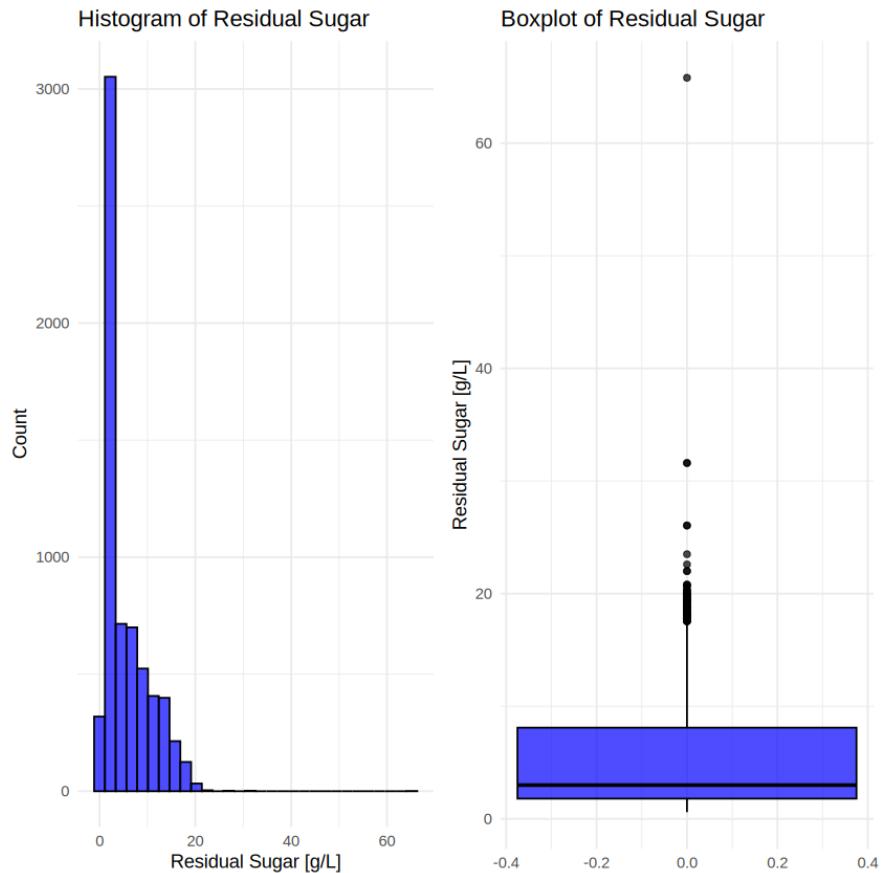


Hình 3.33: Phân phối Lượng muối sunphat trong rượu.

Nhận xét:

- Hầu hết các loại rượu vang có nồng độ sulfat khoảng 0,5 g/L. Có thể thấy hai nhóm ngoại lệ nhỏ khoảng 1,6 và 1,9 g/L trong biểu đồ hộp.

Khảo sát lượng đường còn lại sau khi lên men trong rượu

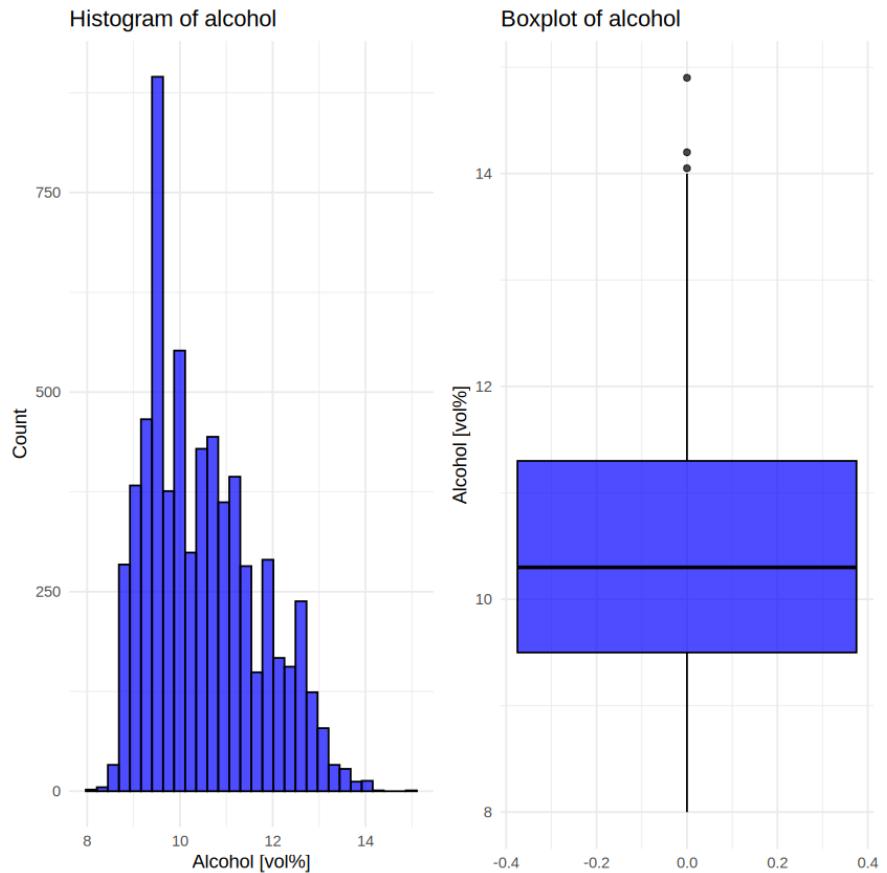


Hình 3.34: Phân phối lượng đường còn lại sau khi lên men trong rượu.

Nhận xét:

- Nhìn chung, các loại rượu vang trong tập dữ liệu có vẻ có nồng độ đường dư thấp. Độ lệch dương di chuyển giá trị trung bình (5,4) lên trên giá trị trung vị (3,0). Có thể tìm thấy giá trị ngoại lệ cực độ xung quanh 65 g/L đường dư

Khảo sát phần trăm cồn trong rượu

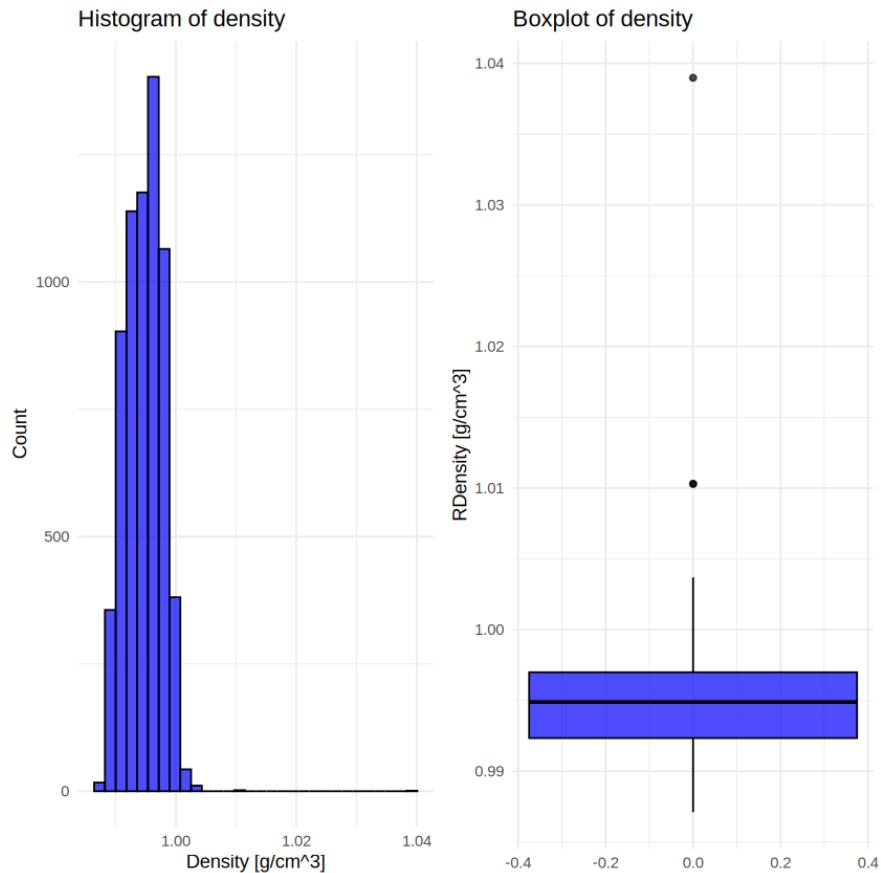


Hình 3.35: Phân phối phần trăm cồn trong rượu.

Nhận xét:

- Hàm lượng cồn của rượu vang trong tập dữ liệu dao động từ 8 đến 15 vol%. Giá trị trung bình nằm trong khoảng 10 vol. Phân phối khá rộng và cho thấy độ lệch dương (lệch phải).

Khảo sát mật độ trong rượu

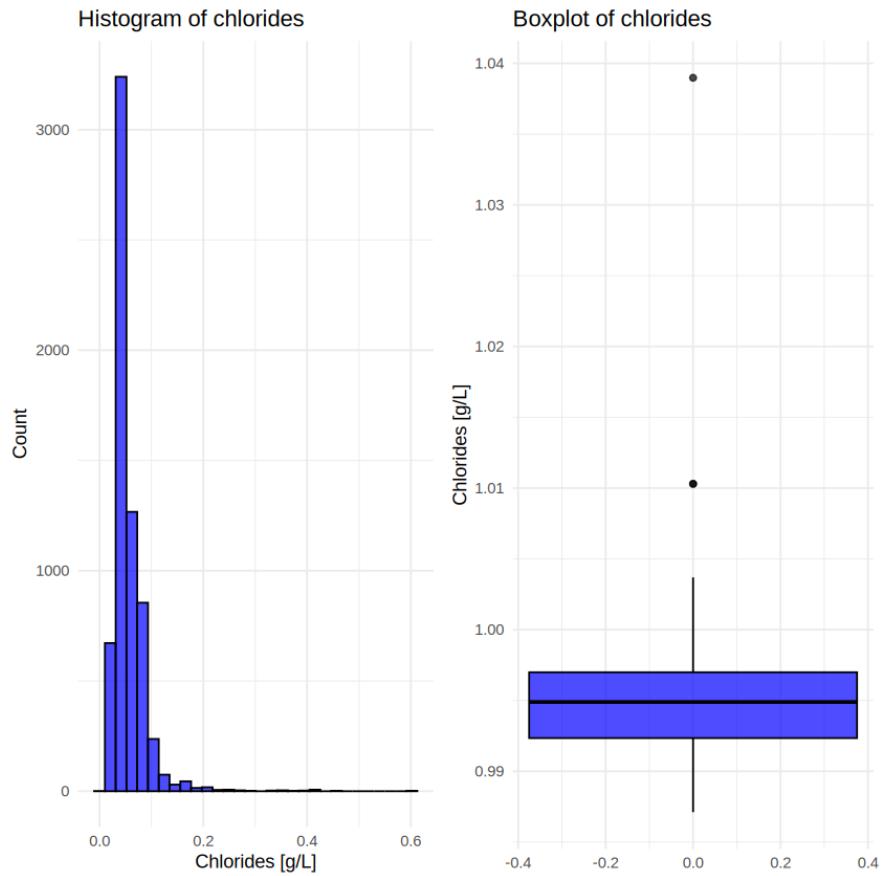


Hình 3.36: Phân phối mật độ rượu.

Nhận xét:

- Tham số mật độ cho thấy sự phân bố rất hẹp với sự thay đổi thấp. Người ta có thể thấy một vài giá trị ngoại lệ trong khoảng 1,01 và 1,04 g/cm³ nhưng hầu hết các loại rượu vang có mật độ trong khoảng 0,99 và 1,00 g/cm³.

Khảo sát lượng muối trong rượu



Hình 3.37: Phân phối lượng muối rượu.

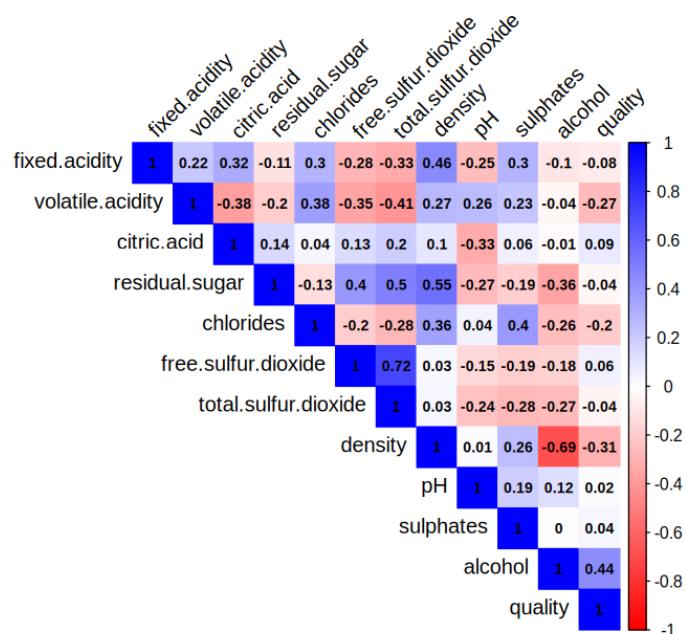
Nhận xét:

- Biểu đồ histogram cho thấy nồng độ clo trong tập dữ liệu có hai đỉnh chính riêng biệt. Nồng độ clo thường gặp nhất có thể được tìm thấy ở khoảng 0,04 g/L. Đỉnh thứ hai xuất hiện ở khoảng 0,08 g/L. Phân phối có đuôi rất dài theo hướng tích cực với các giá trị ngoại lệ lên đến 0,6 g/L.

Tiểu kết phần phân tích đơn biến. Một số nhận xét chính:

- Hầu hết các loại rượu vang đều có xếp hạng chất lượng là 6. Không có loại rượu nào đạt điểm tối đa là 10.
- Độ axit được đo bằng các thành phần cố định và dễ bay hơi. Hầu hết các loại rượu vang đều có độ pH là 3,2.
- Nồng độ lưu huỳnh dioxit thay đổi rất nhiều trong các loại rượu vang được nghiên cứu.
- Các loại rượu vang có hàm lượng cồn dao động từ 8 đến 15 vol%. Nồng độ clorua cho thấy sự phân bố bimodal với các đỉnh ở 0,04 và 0,08 g/L.

Phân tích ma trận tương quan



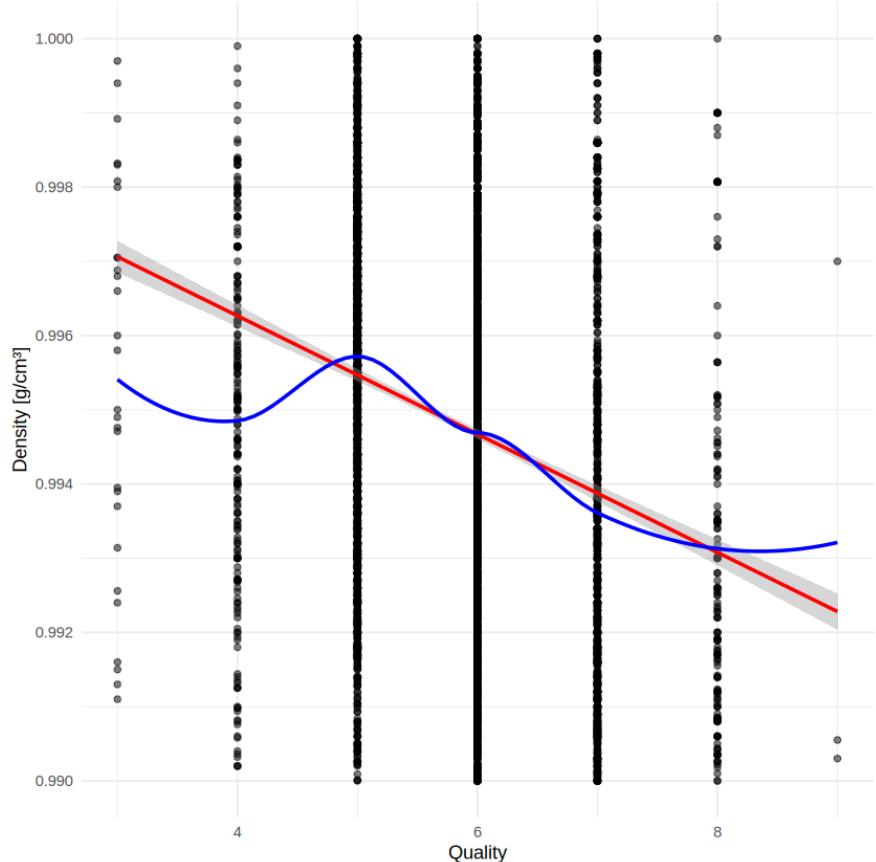
Hình 3.38: Ma trận tương quan giữa các biến trong tập dữ liệu về rượu.

Nhận xét:

- Chọn ngưỡng là 0.3, ta thấy:
 - Nồng độ cồn (alcohol) có ảnh hưởng (thuận) đến chất lượng rượu (chỉ số tương quan 0.436)
 - Các biến ‘residual.sugar’ và ‘density’ có tương quan thuận cao 0.83
 - Mật độ trong rượu (‘density’) có ảnh hưởng (nghịch) đến chất lượng của rượu (chỉ số tương quan -0.307)
 - Các biến ‘alcohol’ và ‘density’ có tương quan nghịch cao -0.78

Phân tích ảnh hưởng của các biến đối với chất lượng rượu

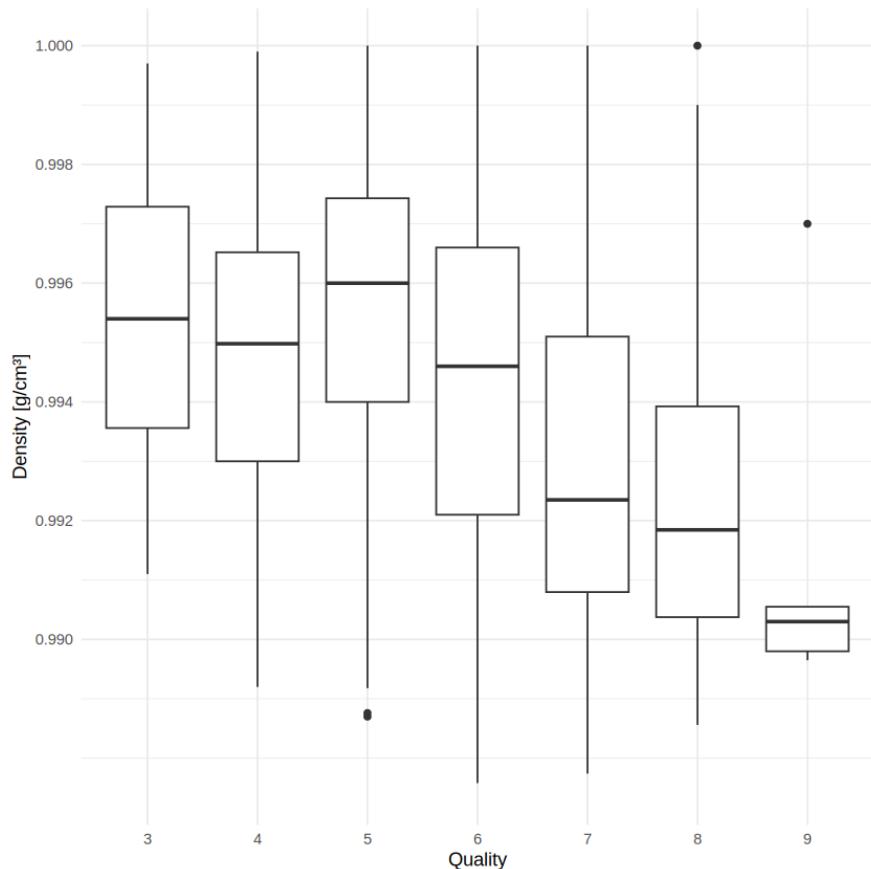
Khảo sát mối quan hệ giữa Density và Quality



Hình 3.39: Mối quan hệ giữa .Density và Quality

Nhận xét:

- Biểu đồ trên cho thấy mật độ rượu được nhóm theo xếp hạng chất lượng của chúng.
- Các giá trị ngoại lệ có thể bị bỏ qua do scale trục y.
- Đường màu xanh lam kết nối các giá trị trung bình của các nhóm chất lượng khác nhau trong khi một đường xu hướng tuyến tính được thêm vào màu đỏ. Chúng ta có thể quan sát thấy xu hướng tiêu cực giữa mật độ và chất lượng nhưng vì chúng ta có các biến thể mật độ lớn trong các nhóm chất lượng khác nhau nên tôi không mong đợi mật độ là một biến dự báo tốt cho chất lượng rượu.

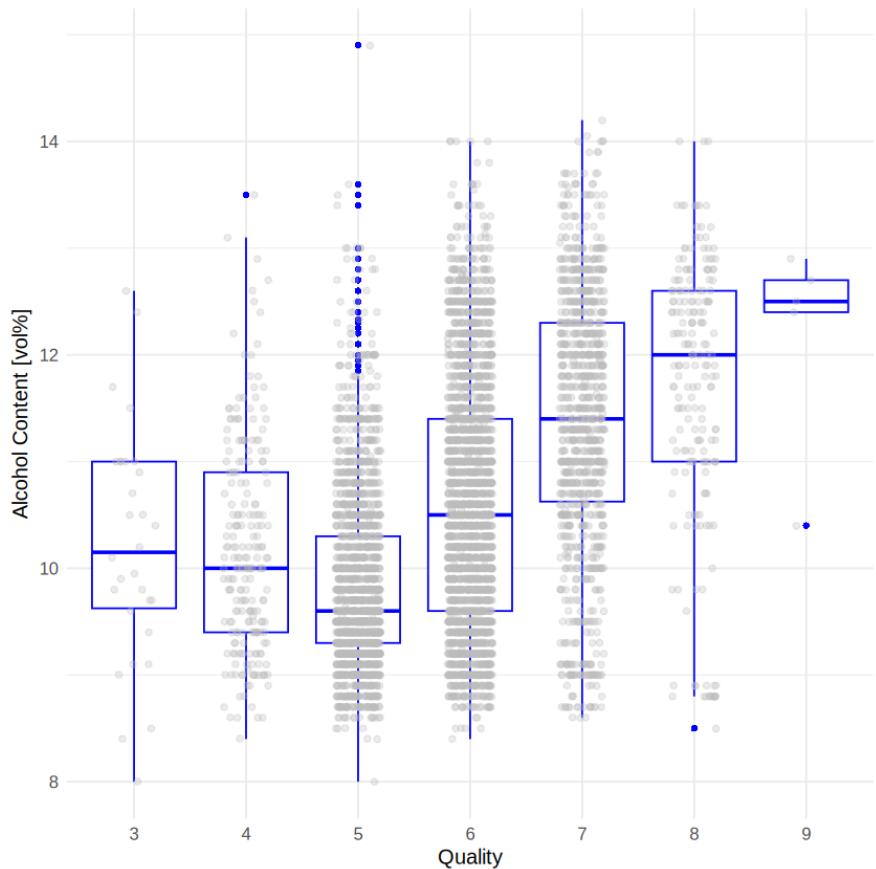


Hình 3.40: Biểu đồ boxplot về mối quan hệ giữa Density và Quality

Nhận xét:

- Chúng ta có được một số nhận định tương tự khi hình dung mối quan hệ giữa chất lượng và mật độ bằng biểu đồ hộp.
- Rượu có mật độ thấp hơn có xu hướng có chất lượng tốt hơn nhưng mật độ thay đổi trong các cửa sổ tương tự trong tất cả các nhóm chất lượng.

Khảo sát mối quan hệ giữa Alcohol và Quality

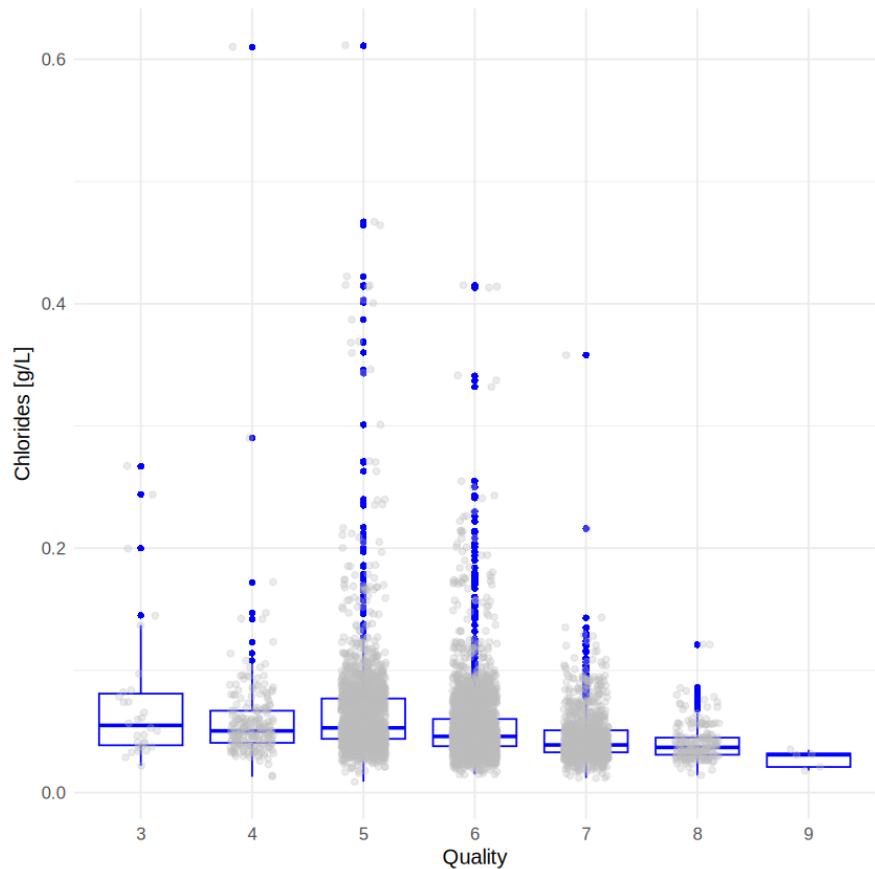


Hình 3.41: Mối quan hệ giữa Alcoholy và Quality

Nhận xét:

- Biểu đồ hộp cho thấy rượu vang có chất lượng cao hơn có vẻ có nồng độ cồn cao hơn. Nhưng mối quan hệ này có vẻ không đáng kể vì các hộp rất rộng và chồng chéo lên nhau đối với các loại khác nhau.

Khảo sát mối quan hệ giữa Chlorides và Quality

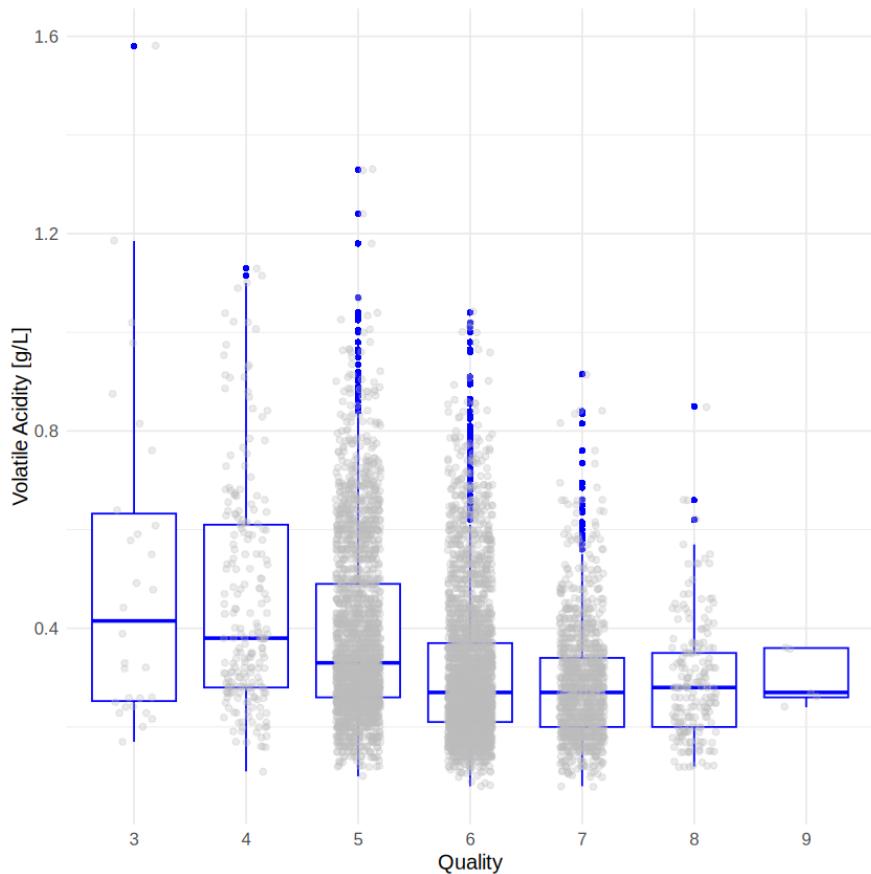


Hình 3.42: Mối quan hệ giữa Chlorides và Quality

Nhận xét:

- Rượu vang có nồng độ clorua thấp hơn có xu hướng có chất lượng tốt hơn nhưng hiệu ứng có vẻ rất yếu. Các khối hộp rỗng và ta có thể thấy rất nhiều ngoại lệ đối với rượu vang chất lượng trung bình.

Khảo sát mối quan hệ giữa Volatile Acidity và Quality

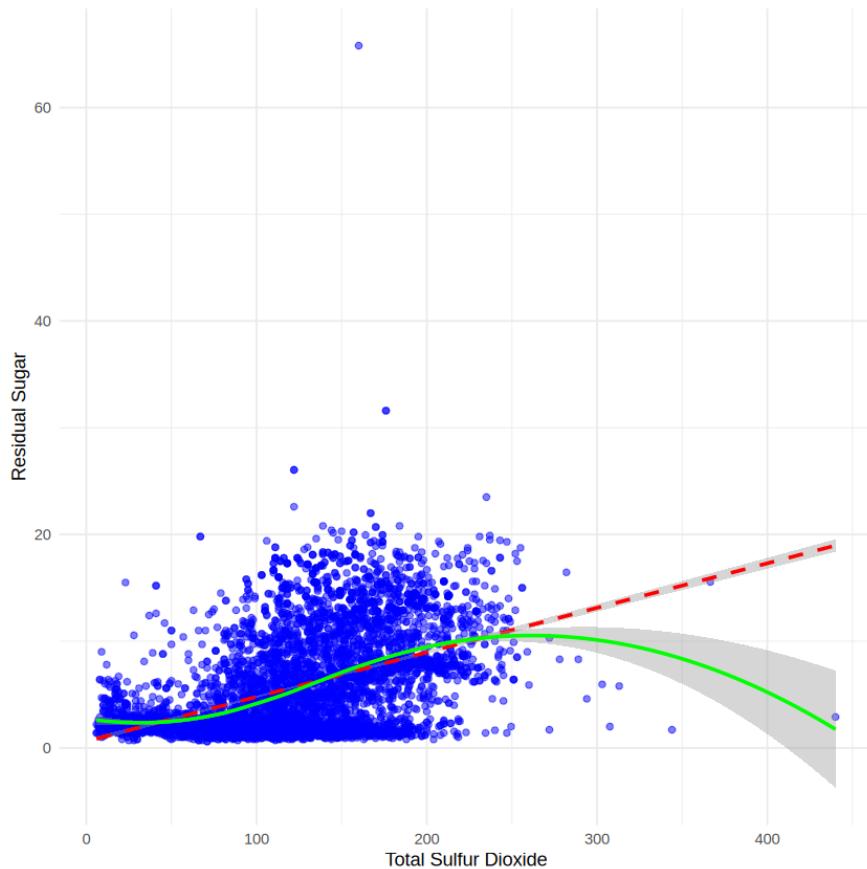


Hình 3.43: Mối quan hệ giữa Volatile Acidity và Quality

Nhận xét:

- Một lần nữa chúng ta chỉ có thể thấy mối tương quan âm rất yếu trong việc hình dung nồng độ axit axetic so với chất lượng rượu.

Khảo sát mối quan hệ giữa tổng lượng SO₂ và lượng đường còn lại sau khi lên men

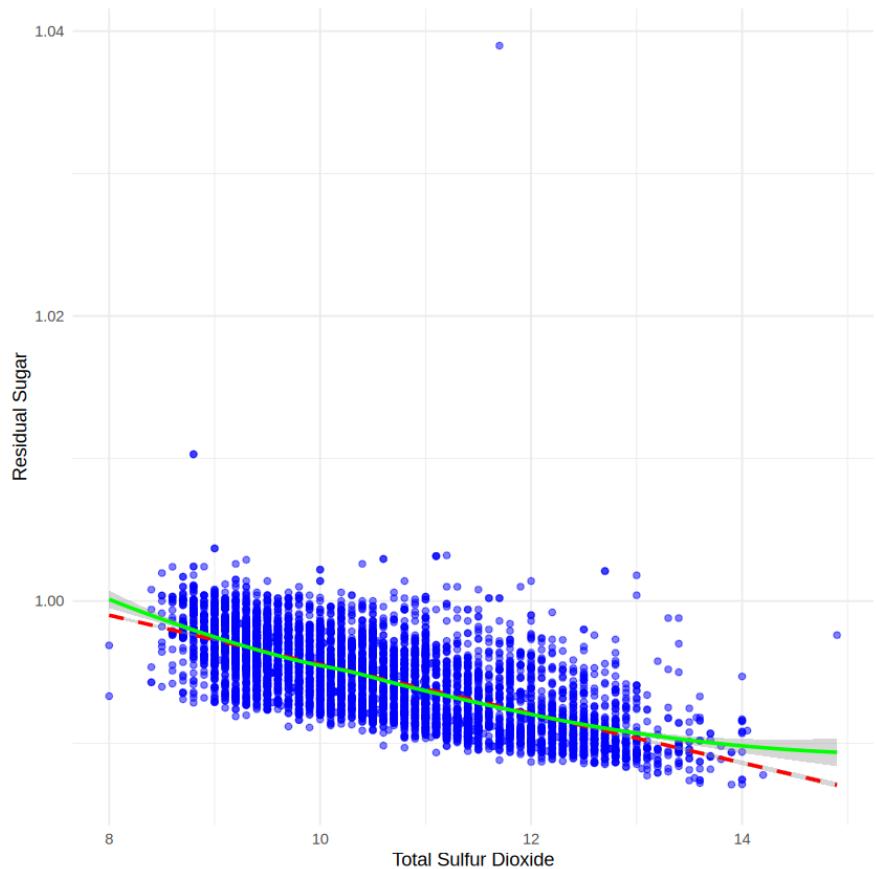


Hình 3.44: Mối quan hệ giữa tổng lượng SO₂ và lượng đường còn lại sau khi lên men

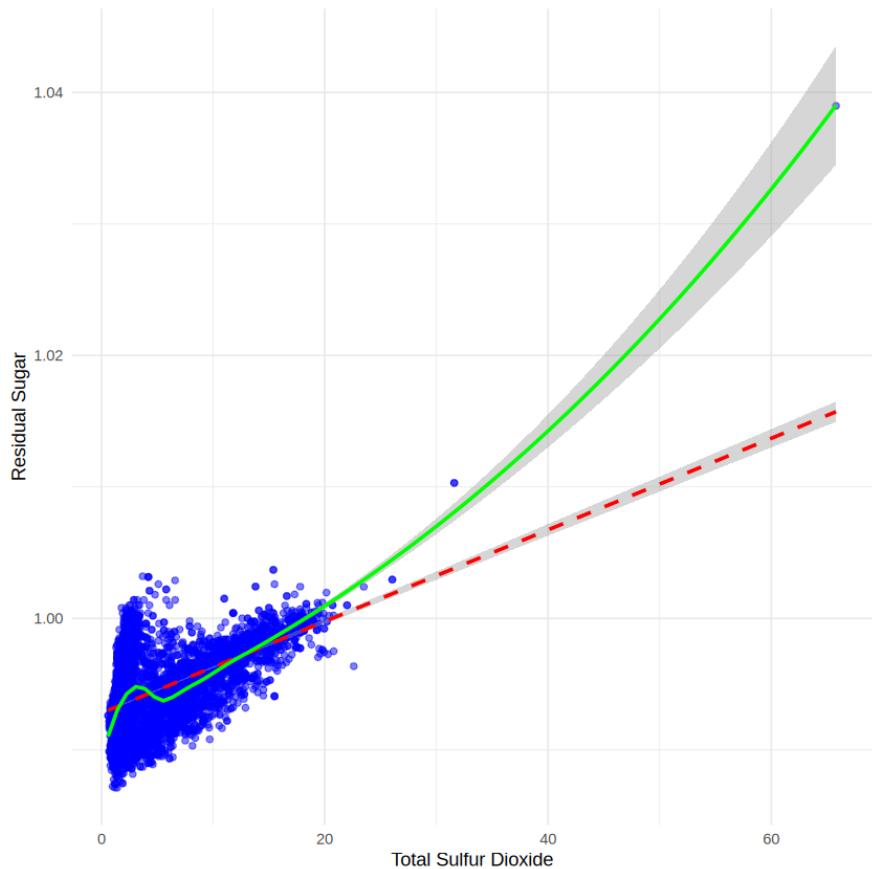
Nhận xét:

- Đối với phần lớn rượu vang, nồng độ lưu huỳnh dioxit tổng thể đường như không bị ảnh hưởng bởi lượng đường còn lại. Chúng ta có thể thấy rằng rượu vang trải dài toàn bộ phạm vi nồng độ lưu huỳnh dioxit đối với mức đường còn lại thấp. Nhưng có vẻ như rượu vang có lượng đường còn lại cao cũng thường có nồng độ lưu huỳnh dioxit tổng thể cao.

Khảo sát tác động của nồng độ cồn, đường dư và lượng đường đến mật độ rượu



Hình 3.45: Mối quan hệ giữa nồng độ cồn và lượng đường đến mật độ rượu



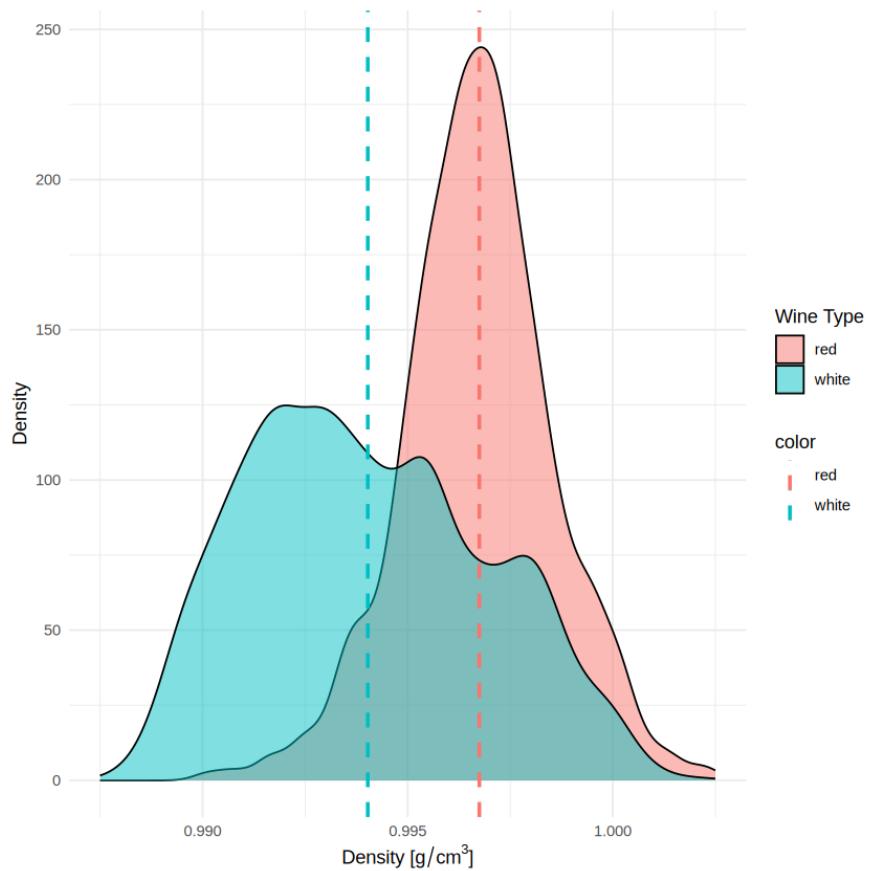
Hình 3.46: Mối quan hệ giữa đường dư và lượng đường đến mật độ rượu

Nhận xét:

- Nồng độ cồn cũng như nồng độ đường còn lại cho thấy ảnh hưởng dự kiến đến mật độ rượu.
- Nồng độ cồn cao hơn làm giảm mật độ rượu trong khi lượng đường còn lại nhiều hơn làm tăng mật độ.
- Đường có mật độ cao hơn nước và do đó làm tăng mật độ của hỗn hợp trong khi rượu thì ngược lại.

Phân tích dựa trên màu sắc của rượu

Phân tích mối quan hệ giữa màu sắc và mật độ rượu

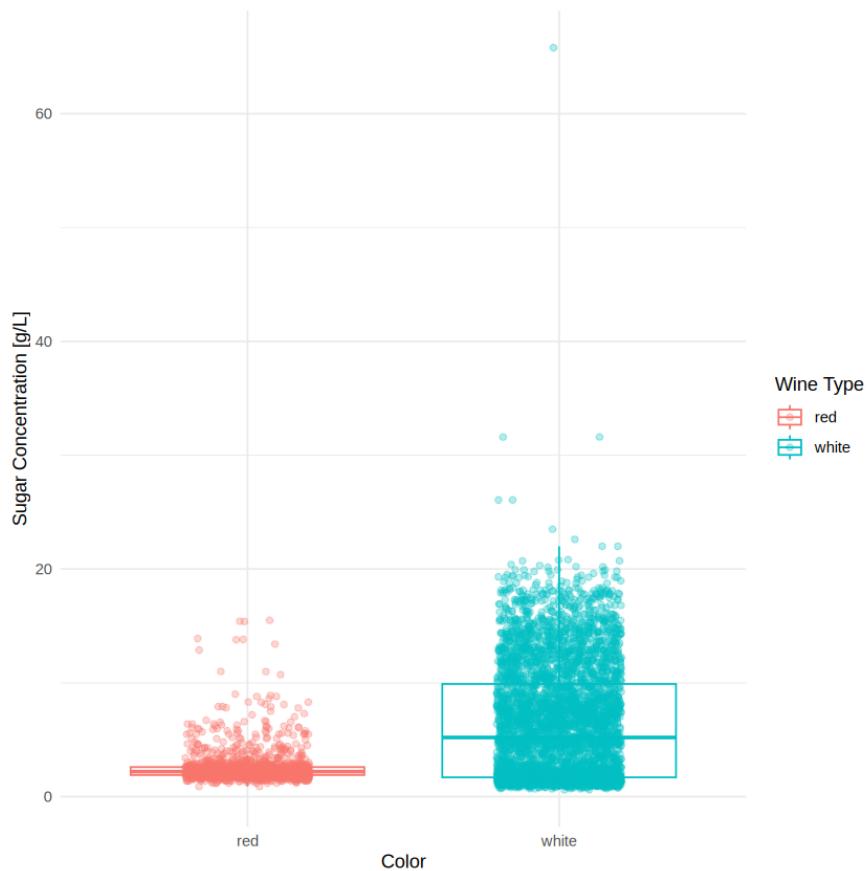


Hình 3.47: Mối quan hệ giữa màu sắc và mật độ rượu

Nhận xét:

- Trong biểu đồ phân bô mật độ ở trên, ta có thể thấy rằng rượu vang trắng và rượu vang đỏ có mật độ khác nhau. Rượu vang đỏ có sự phân bô hẹp với giá trị trung bình khoảng 0,997 g/cm³. Ngược lại, rượu vang trắng có sự thay đổi mật độ rộng hơn nhiều nhưng trung bình của chúng nằm dưới giá trị trung bình của rượu vang đỏ khoảng 0,993 g/cm³.

Phân tích mối quan hệ giữa màu sắc và lượng đường còn lại sau khi lên men

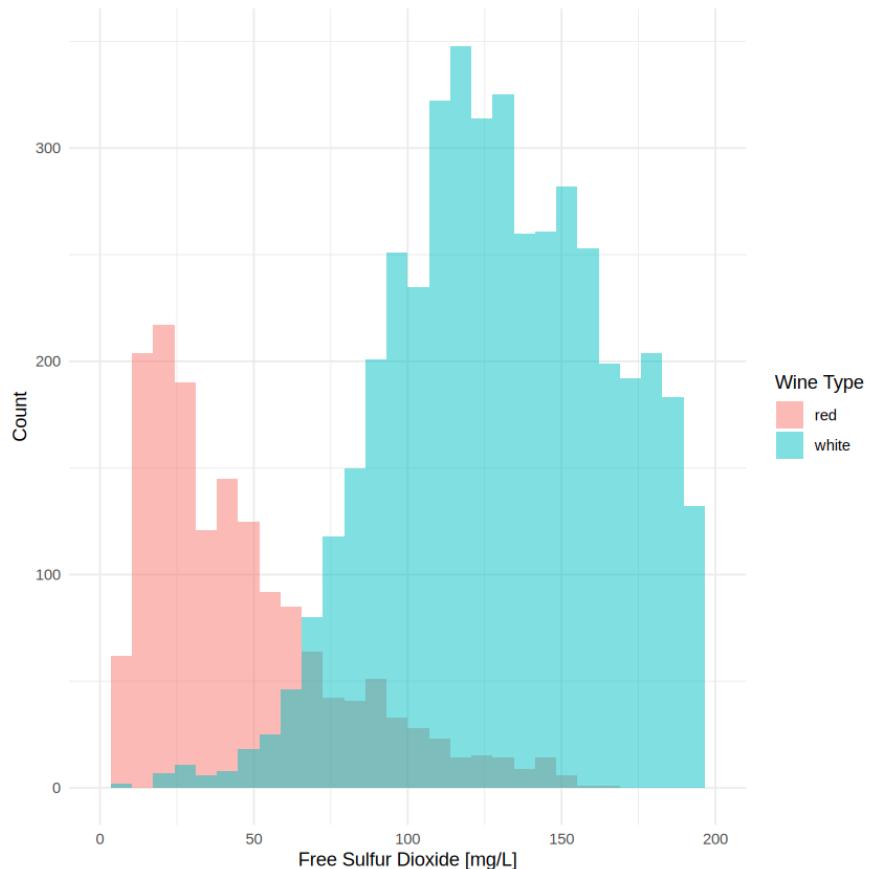


Hình 3.48: Mối quan hệ giữa màu sắc và lượng đường còn lại sau khi lên men

Nhận xét:

- Nồng độ đường còn lại thấp hơn ở rượu vang đỏ so với rượu vang trắng và sự phân bố của chúng hẹp hơn.

Phân tích mối quan hệ giữa màu sắc và tổng lượng lưu huỳnh trong rượu

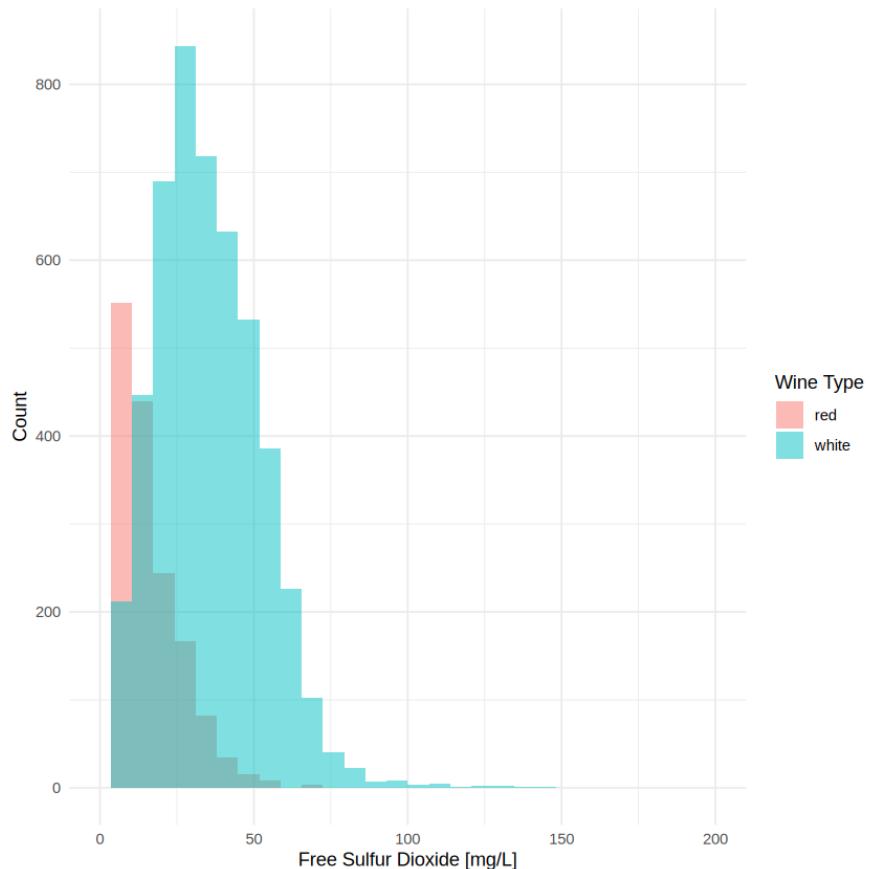


Hình 3.49: Mối quan hệ giữa màu sắc và tổng lượng lưu huỳnh trong rượu

Nhận xét:

- Biểu đồ ở trên cho thấy mối quan hệ giữa màu rượu vang và nồng độ lưu huỳnh dioxit tổng thể. Hầu hết rượu vang đỏ có nồng độ lưu huỳnh dioxit tổng thể thấp hơn khi rượu vang trắng cho thấy sự phân bố đối xứng xung quanh giá trị trung bình cao hơn là 130 g/L.

Phân tích mối quan hệ giữa màu sắc và lượng lưu huỳnh tự do trong rượu

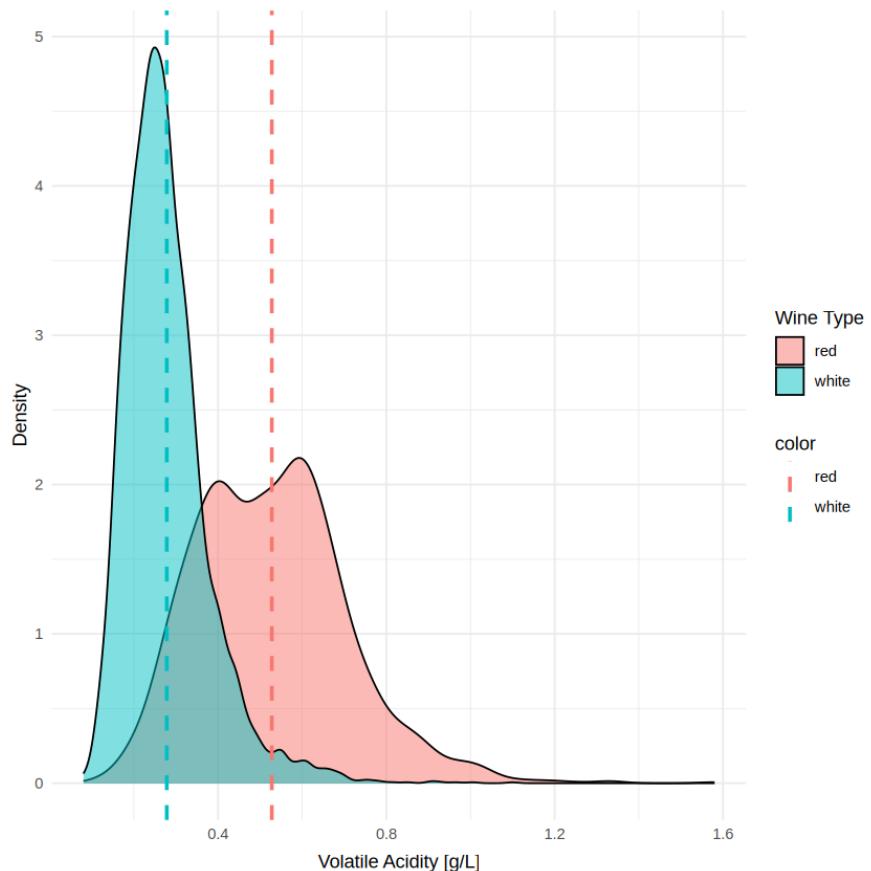


Hình 3.50: Mối quan hệ giữa màu sắc và lượng lưu huỳnh tự do trong rượu

Nhận xét:

- Khi xem xét sự khác biệt giữa lưu huỳnh dioxit tổng và tự do, chúng ta có thể khẳng định rõ ràng rằng lưu huỳnh dioxit cố định là nguyên nhân tạo nên sự khác biệt về màu sắc của rượu vang.

Phân tích mối quan hệ giữa màu sắc và tính chua của rượu



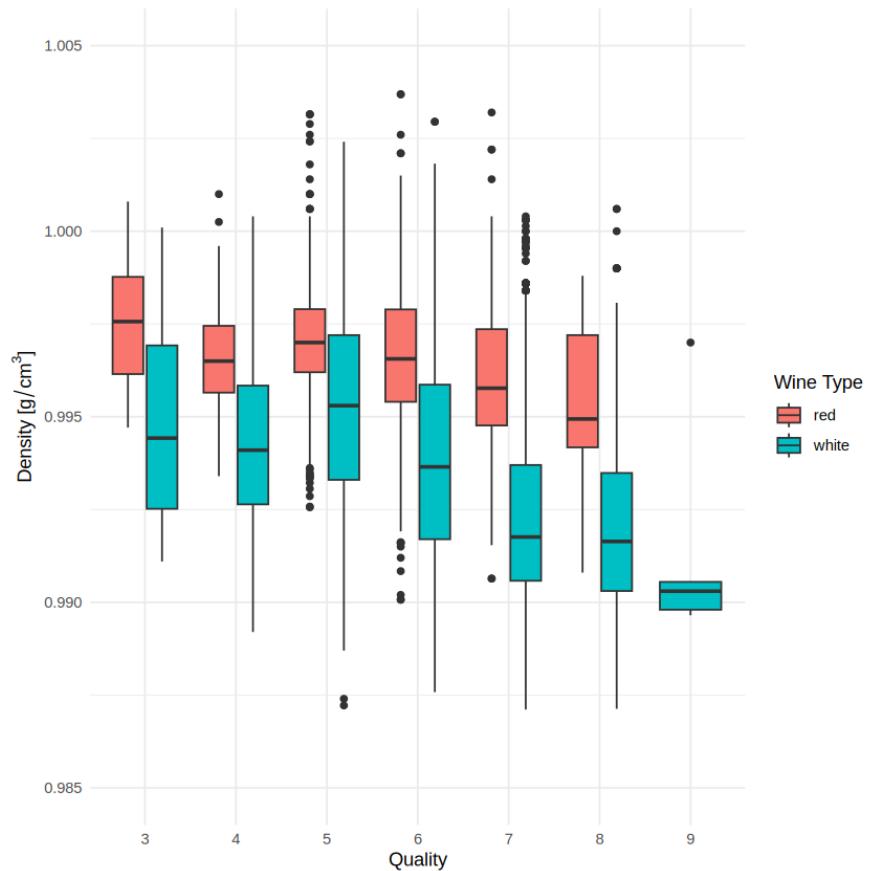
Hình 3.51: Mối quan hệ giữa màu sắc và tính chua của rượu

Nhận xét:

- Sự phân bố của độ axit dễ bay hơi cho thấy đỉnh chính của nó ở nồng độ thấp hơn đối với rượu vang trắng so với rượu vang đỏ. Trong khi đường cong đối với rượu vang trắng hẹp với độ lệch dương, đường cong mật độ rượu vang đỏ rộng và cho thấy dấu hiệu của phân phối hai đỉnh.

Phân tích tương quan giữa các biến dựa trên màu sắc

Khảo sát tương quan giữa mật độ và chất lượng rượu dựa trên màu sắc

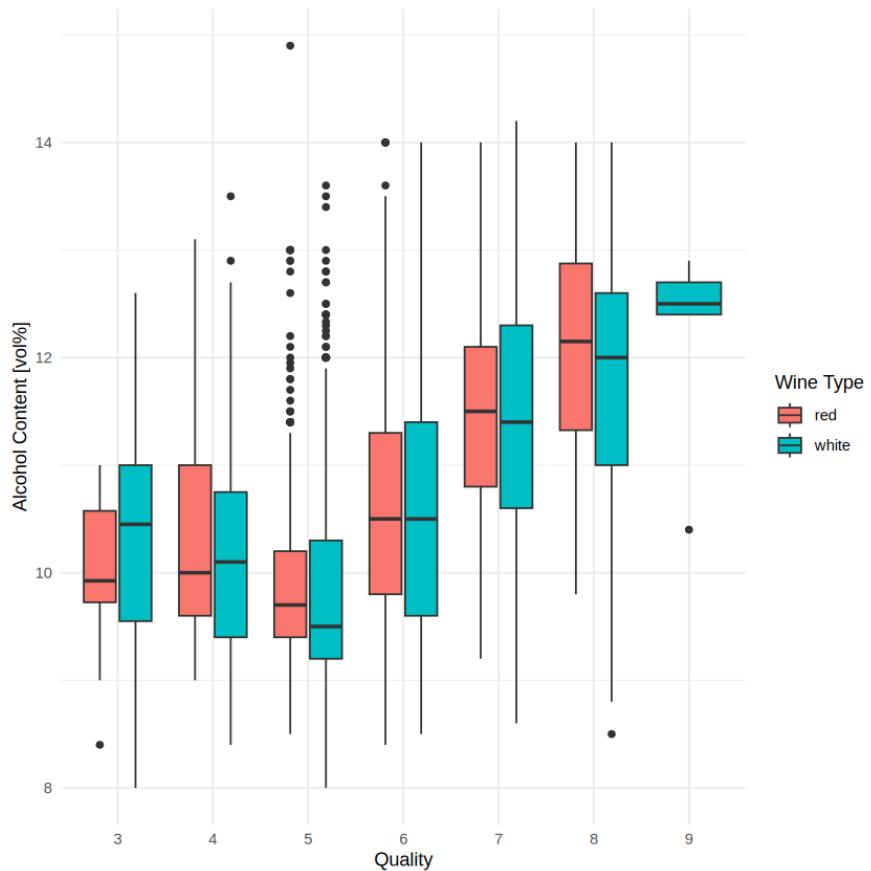


Hình 3.52: Mối quan hệ giữa mật độ và chất lượng rượu dựa trên màu sắc

Nhận xét:

- Đối với cả hai nhóm rượu, chất lượng có tương quan nghịch với mật độ. Nhưng tác dụng có vẻ mạnh hơn một chút đối với rượu vang trắng.

Khảo sát tương quan giữa nồng độ cồn và chất lượng rượu dựa trên màu sắc

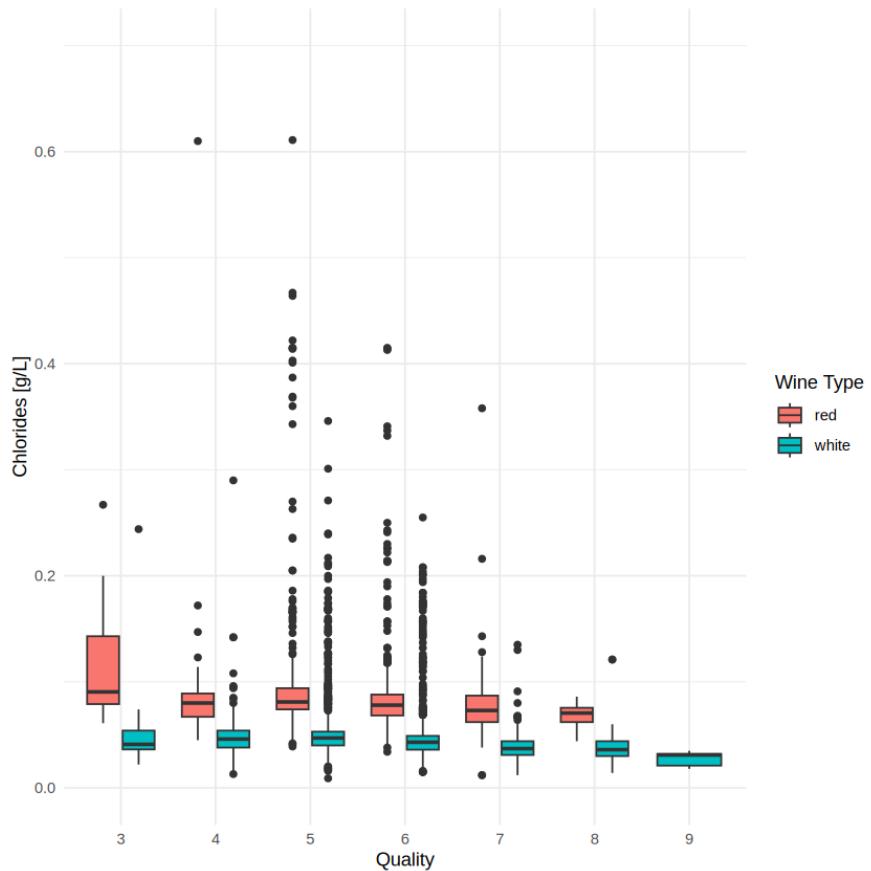


Hình 3.53: Mối quan hệ giữa nồng độ cồn và chất lượng rượu dựa trên màu sắc

Nhận xét:

- Ngoài ra đối với nồng độ cồn, ta thấy xu hướng tương tự đối với cả hai màu rượu.

Khảo sát tương quan giữa lượng muối và chất lượng rượu dựa trên màu sắc

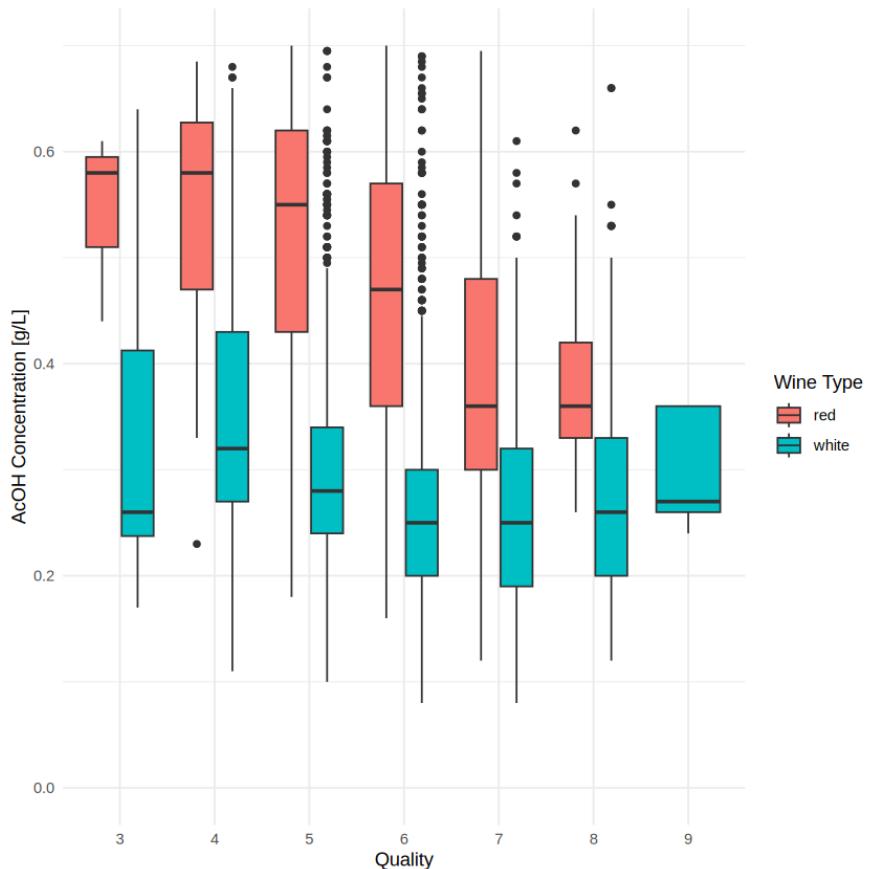


Hình 3.54: Mối quan hệ giữa lượng muối và chất lượng rượu dựa trên màu sắc

Nhận xét:

- Rượu vang trắng nhìn chung có nồng độ clorua thấp hơn rượu vang đỏ. Nhưng cả hai nhóm đều cho thấy xu hướng tương quan nghịch giữa nồng độ clorua và đánh giá chất lượng nhưng tác động rất yếu và có rất nhiều ngoại lệ đối với rượu vang chất lượng trung bình.

Khảo sát tương quan giữa độ chua và chất lượng rượu dựa trên màu sắc



Hình 3.55: Mối quan hệ giữa độ chua và chất lượng rượu dựa trên màu sắc

Nhận xét:

- Nồng độ axit axetic có ảnh hưởng đến chất lượng rượu vang đỏ
- Đối với rượu vang trắng, ta thấy có khá nhiều ngoại lai ở chất lượng 8

Khảo sát đa công tuyến

Bước 1: Tính toán chỉ số VIF

	fixed.acidity	volatile.acidity	citric.acid
1	5.048348	2.168159	1.622151
2	residual.sugar	chlorides	free.sulfur.dioxide
3	9.634653	1.659342	2.235693
4	total.sulfur.dioxide	density	pH
5	4.045899	22.337223	2.563776
6	sulphates	alcohol	color
7	1.555807	5.616857	7.224467
8			

Nhận xét:

- Ta có chọn ngưỡng bằng 3

Bước 2: Loại bỏ các biến dựa trên VIF nếu vượt quá ngưỡng

```
1 fixed.acidity      volatile.acidity      citric.acid
2             1.783515                  1.703665          1.608022
3 residual.sugar      chlorides      free.sulfur.dioxide
4             1.511206                  1.564130          2.135374
5 total.sulfur.dioxide      pH      sulphates
6             2.843819                  1.415649          1.347969
7 alcohol
8             1.410019
9
10 Call :
11 lm(formula = quality ~ fixed.acidity + volatile.acidity +
12   citric.acid +
13   residual.sugar + chlorides + free.sulfur.dioxide + total.
14   sulfur.dioxide +
15   pH + sulphates + alcohol, data = wine_quality)
16
17 Residuals:
18
19 Coefficients:
20
21 (Intercept)      Estimate Std. Error t value Pr(>|t|)    
22 fixed.acidity      1.9132915  0.2755174  6.944 4.17e-12 ***
23 volatile.acidity     -0.0114495  0.0094124  1.216  0.2239    
24 citric.acid       -1.4523016  0.0724401 -20.048 < 2e-16 ***
25 residual.sugar      0.0227933  0.0023608  9.655 < 2e-16 ***
26 chlorides        -0.7908671  0.3261855 -2.425  0.0154 *  
27 free.sulfur.dioxide  0.0059939  0.0007523  7.968 1.89e-15 ***
28 total.sulfur.dioxide -0.0022574  0.0002726 -8.281 < 2e-16 ***
29 pH                 0.1672385  0.0676144  2.473  0.0134 *  
30 sulphates         0.6460948  0.0712907  9.063 < 2e-16 ***
31 alcohol           0.3306436  0.0090968 36.347 < 2e-16 ***
32 ---
```

```

33 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
34
35 Residual standard error: 0.7364 on 6486 degrees of freedom
36 Multiple R-squared:  0.2899 ,   Adjusted R-squared:  0.2888
37 F-statistic: 264.8 on 10 and 6486 DF,  p-value: < 2.2e-16

```

Khảo sát ngoại lai

Ta sử dụng IQR để tìm các điểm ngoại lai và cực ngoại lai:

- Tổng số ngoại lai: 1657
- Tổng số cực ngoại lai: 304

Trong bài toán này, ta sẽ loại bỏ các điểm cực ngoại lai

Chuẩn hóa và phân chia tập dữ liệu

Ta sử dụng box-cox transform và sau đó phân chia tập dữ liệu thành 2 phần: train (80%) và test (20%).

Xây dựng mô hình

```

1 # Xây dựng mô hình đầy đủ
2 full.lm <- lm(quality ~ ., data = train)
3 print(summary(full.lm))

4

5 # Mô hình chặn dưới
6 model.lb <- lm(quality ~ 1, data = train)

7

8 # Mô hình chặn trên
9 model.up <- full.lm

10

11 step(full.lm, scope = list(lower = model.lb, upper = model.up),
      direction = "both", trace = FALSE)

```

Kết quả:

```

1 lm(formula = quality ~ fixed.acidity + volatile.acidity +
     citric.acid +
2     residual.sugar + chlorides + free.sulfur.dioxide +
     sulphates +

```

```

3     alcohol, data = train)

4

5 Coefficients:
6             (Intercept)      fixed.acidity    volatile.acidity
7             -4.973e-01       -6.243e-02      -1.378e-04
8             citric.acid      residual.sugar   chlorides
9             4.438e-03        5.123e-03      -1.505e-06
10            free.sulfur.dioxide sulphates      alcohol
11            2.354e-02        5.480e-04      2.013e+00

```

```

1 wqr_models <- regsubsets(quality ~ volatile.acidity + chlorides
2                           + density + pH + sulphates + alcohol, data = train)
3 summary.wqr <- summary(wqr_models)

```

Ta lựa chọn mô hình tốt nhất dựa trên BIC. Kết quả:

```

1 lm(formula = as.formula(formula_str), data = train)
2
3 Residuals:
4      Min       1Q   Median       3Q      Max
5 -0.042990 -0.002084  0.000506  0.002979  0.013610
6
7 Coefficients:
8                 Estimate Std. Error t value Pr(>|t|)
9 (Intercept) -5.274e-01  3.736e-02 -14.117 < 2e-16 ***
10 volatile.acidity -1.423e-04  1.068e-05 -13.321 < 2e-16 ***
11 citric.acid 3.596e-03  1.016e-03   3.538 0.000407 ***
12 residual.sugar 5.113e-03  5.168e-04   9.894 < 2e-16 ***
13 chlorides -1.691e-06  3.393e-07  -4.983 6.48e-07 ***
14 free.sulfur.dioxide 2.516e-02  5.535e-03   4.546 5.59e-06 ***
15 sulphates 5.368e-04  6.610e-05   8.121 5.78e-16 ***
16 alcohol 2.011e+00  7.478e-02   26.889 < 2e-16 ***
17 ---
18 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
19
20 Residual standard error: 0.004731 on 4946 degrees of freedom
21 Multiple R-squared:  0.2087,    Adjusted R-squared:  0.2076
22 F-statistic: 186.4 on 7 and 4946 DF,  p-value: < 2.2e-16

```

Dánh giá hiệu suất và dự đoán kết quả

3.1.6. Kết luận

3.2. Phân tích chất lượng không khí

3.2.1. Giới thiệu chung

Trong lĩnh vực khoa học dữ liệu và mô hình hóa thống kê, việc xử lý dữ liệu cao chiều (high-dimensional data) với đa cộng tuyến giữa các biến dự báo là một thách thức phổ biến. Các kỹ thuật hồi quy truyền thống thường gặp khó khăn trong các tình huống như vậy, dẫn đến ước tính không ổn định và hiệu suất dự báo kém. Để giải quyết các vấn đề này, các kỹ thuật giảm chiều như Hồi quy thành phần chính (Principal Component Regression - PCR) và Hồi quy bình phương tối thiểu riêng phần (Partial Least Squares Regression - PLS) được sử dụng. Trong đồ án này, chúng tôi thực nghiệm nghiên cứu và áp dụng hai kỹ thuật này thông qua bộ dữ liệu thực tế về đánh giá chất lượng không khí.

3.2.2. Phát biểu bài toán

Trong đồ án này, chúng tôi quan tâm đến vấn đề chất lượng không khí mà trong đó chúng tôi quan tâm đến nồng độ chất C₆H₆ (benzene), được ký hiệu trong dữ liệu là C₆H₆(GT) dựa trên ý nghĩa của nó đối với sức khỏe cộng đồng và môi trường môi sinh. Benzen là một chất gây ô nhiễm chính, được phân loại là chất gây ung thư và sự hiện diện của nó trong khí quyển có liên quan chặt chẽ đến nhiều nguy cơ sức khỏe, bao gồm cả tỷ lệ ung thư gia tăng. Ngoài ra, nồng độ benzen đóng vai trò là chỉ số về khí thải từ phương tiện giao thông và công nghiệp, đây là những mối quan tâm chính trong quản lý ô nhiễm đô thị.

Ý nghĩa của đồ án này nằm ở việc phân tích và đưa ra các thông tin hữu ích có giá trị từ khảo sát dữ liệu chất lượng không khí, từ đó giúp người quản lý có thể đưa ra những chiến lược phù hợp nhằm cải thiện chất lượng không khí.

3.2.3. Giới thiệu về dữ liệu

Bộ dữ liệu này, The Air Quality Dataset, được lấy từ UCI Machine Learning Repository, chứa các chỉ số đo đạc các chất gây ô nhiễm không khí và các biến số khí tượng khác nhau được thu thập tại một trạm giám sát của Ý. Bộ dữ liệu bao gồm các phép đo hàng ngày về các chất gây ô nhiễm như ôzôn, nitơ dioxit và cacbon monoxit, cũng như các biến số khí tượng như nhiệt độ, tốc độ gió và độ ẩm. Với hơn 9.000 quan trắc, bộ dữ liệu này cung cấp một nguồn dữ liệu phong phú để phân tích.

Trong đồ án này, biến mục tiêu là một trong những chất gây ô nhiễm, cho phép sử dụng các mô hình hồi quy như PCR và PLS để dự đoán chất lượng không khí dựa trên các yếu tố dự báo sẵn có.

3.2.4. Khám phá và tiền xử lý dữ liệu

Dữ liệu này có 15 cột và 9357 dòng. Dựa trên thông tin của tập dữ liệu, ta thấy mỗi dòng mang ý nghĩa khác nhau, tức là mỗi quan trắc độc lập nhau. Và ta cũng dễ dàng kiểm tra được dữ liệu không có hiện tượng trùng lặp.

Ta có ý nghĩa của từng cột như sau:

Bảng 3.1: Ý nghĩa các cột của dữ liệu chất lượng không khí.

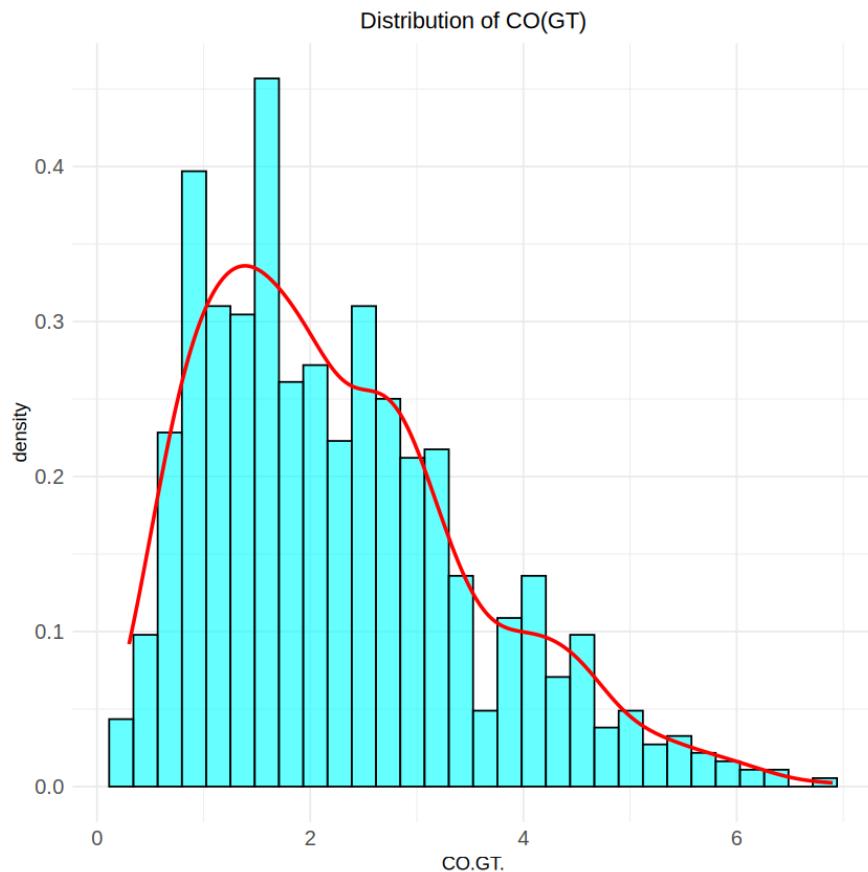
Tên biến	Phân loại	Mô tả	Đơn vị	Missing Values
Date	Date (Ngày tháng năm)	Ngày tháng mà các giá trị độ đo được thu thập.		Không
Time	Categorical (Phân loại)	Thời gian mà các giá trị độ đo được thu thập.		Không
CO(GT)	Integer (Nguyên)	Nồng độ CO (carbon monoxide) trung bình thực tế theo giờ tính bằng mg/m ³ (máy phân tích tham chiếu)	mg/m ³	Không
PT08.S1(CO)	Categorical (Phân loại)	(Trung bình theo giờ) Phản hồi cảm biến (PPM) của cảm biến hồng ngoại không phân tán (NDIR) đối với carbon monoxide.		Không
NMHC(GT)	Integer (Nguyên)	Nồng độ Non Metanic HydroCarbons trong không khí (đơn vị microg/m ³ / trung bình theo giờ/ dựa trên máy phân tích tham chiếu)	microg/m ³	Không
C6H6(GT)	Continuous	Nồng độ trung bình Benzene trong không khí (đơn vị microg/m ³ / trung bình theo giờ/ dựa trên máy phân tích tham chiếu)	microg/m ³	Không

PT08.S2(NMHC)	Categorical (Phân loại)	(Trung bình theo giờ) Phản hồi cảm biến (PPM) của cảm biến hồng ngoại không phân tán (NDIR) đối với Non Metanic HydroCarbons		Không
NOx(GT)	Integer (Nguyên)	Nồng độ trung bình NOx (oxit nitơ) trong không khí (đơn vị PPB/ trung bình theo giờ/ dựa trên máy phân tích tham chiếu)	ppb	Không
PT08.S3(NOx)	Categorical (Phân loại)	(Trung bình theo giờ) Phản hồi cảm biến (PPM) của cảm biến hồng ngoại không phân tán (NDIR) đối với NOx (oxit nitơ)		Không
NO2(GT)	Integer (Nguyên)	Nồng độ trung bình NO2 (nitơ dioxit) trong không khí (đơn vị microg/m ³ / trung bình theo giờ/ dựa trên máy phân tích tham chiếu)	microg/m ³	Không
PT08.S4(NO2)	Categorical (Phân loại)	(Trung bình theo giờ) Phản hồi cảm biến (PPM) của cảm biến hồng ngoại không phân tán (NDIR) đối với NO2 (nitơ dioxit)		Không
PT08.S5(O3)	Categorical (Phân loại)	(Trung bình theo giờ) Phản hồi cảm biến (PPM) của cảm biến hồng ngoại không phân tán (NDIR) đối với O3		Không
T	Continuous (Liên tục)	Nhiệt độ (Thang độ Celsius)	°C	Không
RH	Continuous (Liên tục)	Dộ ẩm tương đối (Relative Humidity)	%	Không

AH	Continuous (Liên tục)	Dộ ẩm tuyệt đối (Absolute Humidity)		Không
----	-----------------------	-------------------------------------	--	-------

3.2.5. Phân tích đơn biến

$CO(GT)$: Carbon monoxide concentration (mg/m^3)

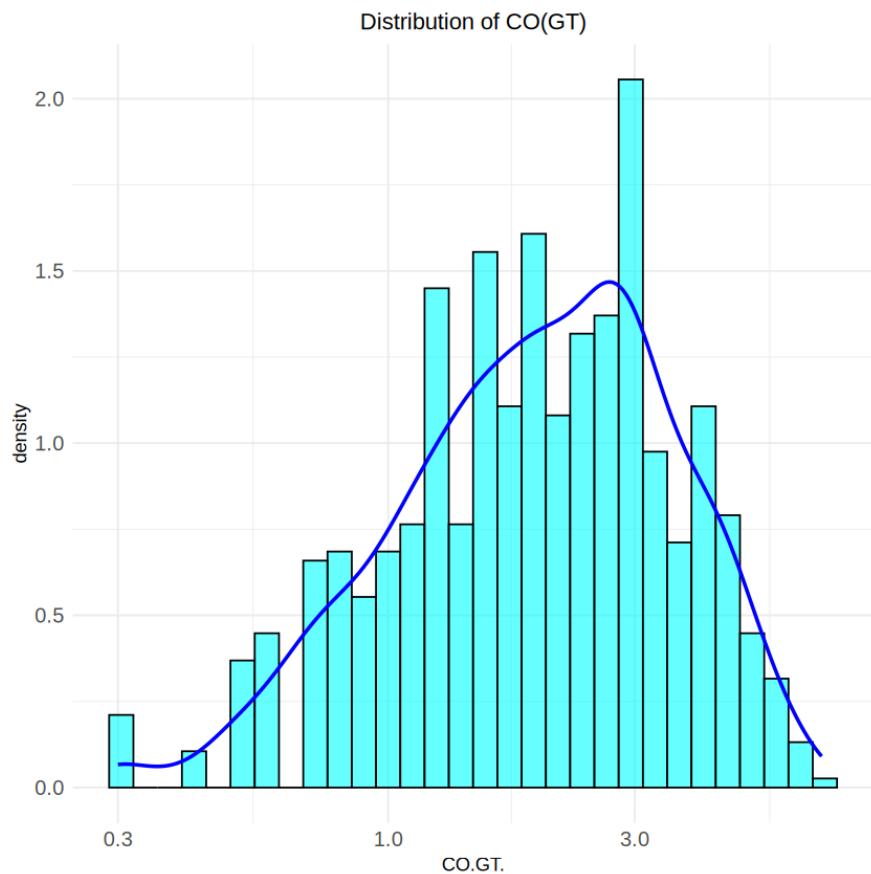


Hình 3.56: Phân phối ban đầu của Carbon monoxide.

Nhận xét:

- Nhìn vào histogram trên, ta thấy phân phối của biến này bị lệch trái (lệch dương).

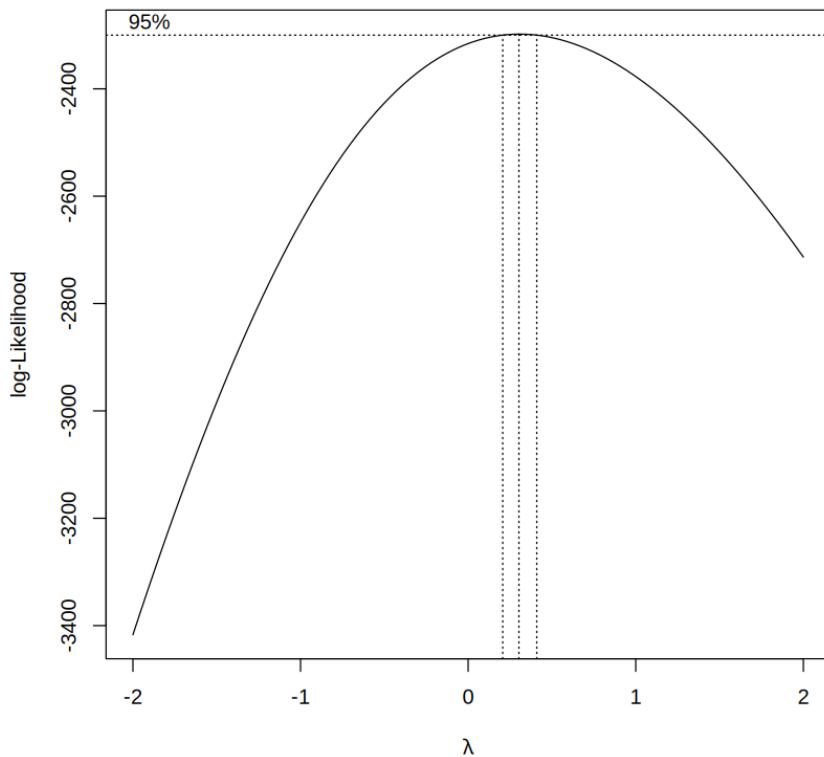
Ta thử sử dụng log-transform nó.



Hình 3.57: Phân phối sau khi log-scale của Carbon monoxide.

Nhận xét:

- Sau khi sử dụng log-transform, hình dạng phân phối tương đối chuẩn hơn.
- Ta có thể thử sử dụng box-cox transform

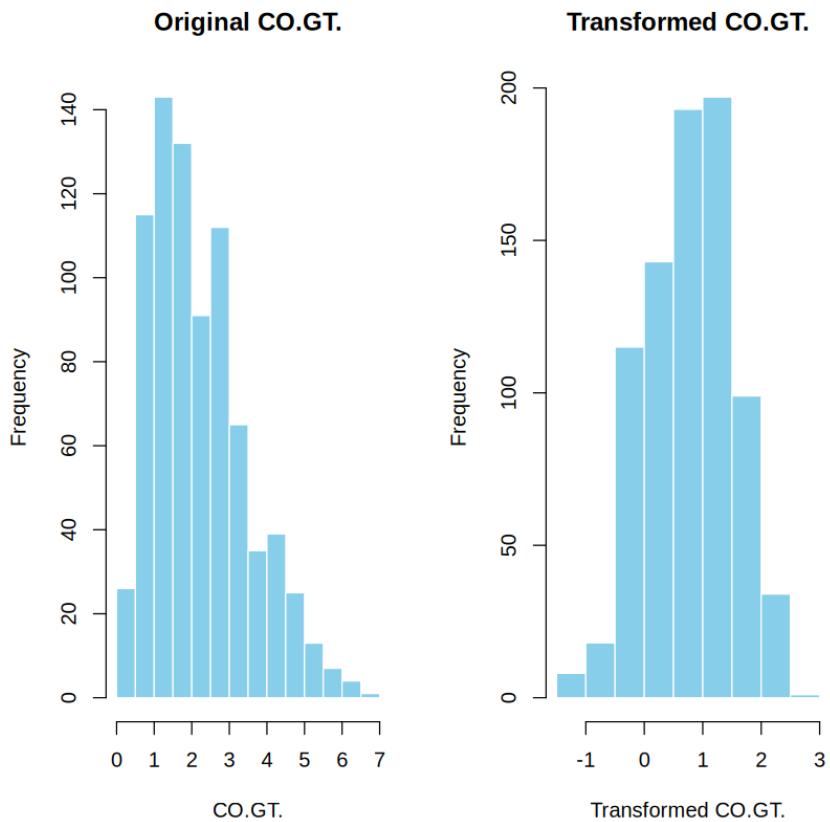


Hình 3.58: Log-likelihood với các giá trị λ của Carbon monoxide.

Nhận xét:

- Dựa trên biểu đồ, ta tìm được giá trị lambda tối ưu với mức ý nghĩa 5% là 0.343

Và ta thực hiện biến đổi dữ liệu với giá trị lambda vừa tìm được.

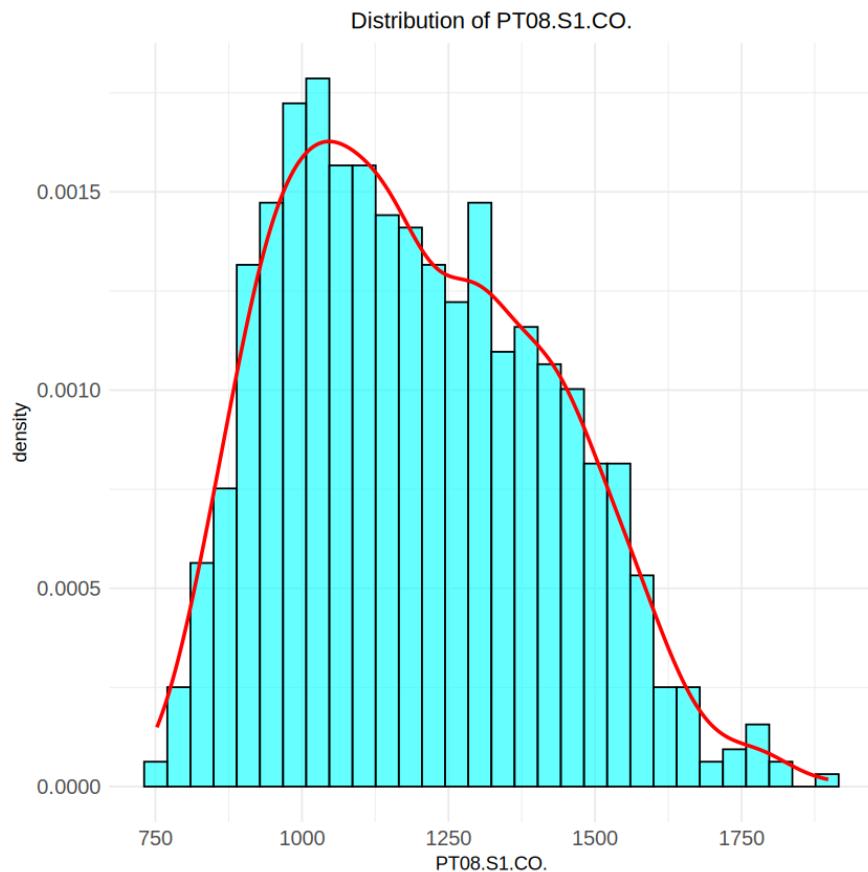


Hình 3.59: Phân phối trước và sau khi biến đổi của Carbon monoxide.

Nhận xét:

- Ta có được giá trị lambda tối ưu là 0.343 và sử dụng giá trị này để biến đổi biến CO(GT). Biểu đồ histogram phía bên dưới thể hiện phân phối của biến này trước và sau khi biến đổi. Dễ dàng thấy được, sau khi biến đổi, biến này đã tương đối chuẩn hơn

PT08.S1(CO): Sensor response for CO

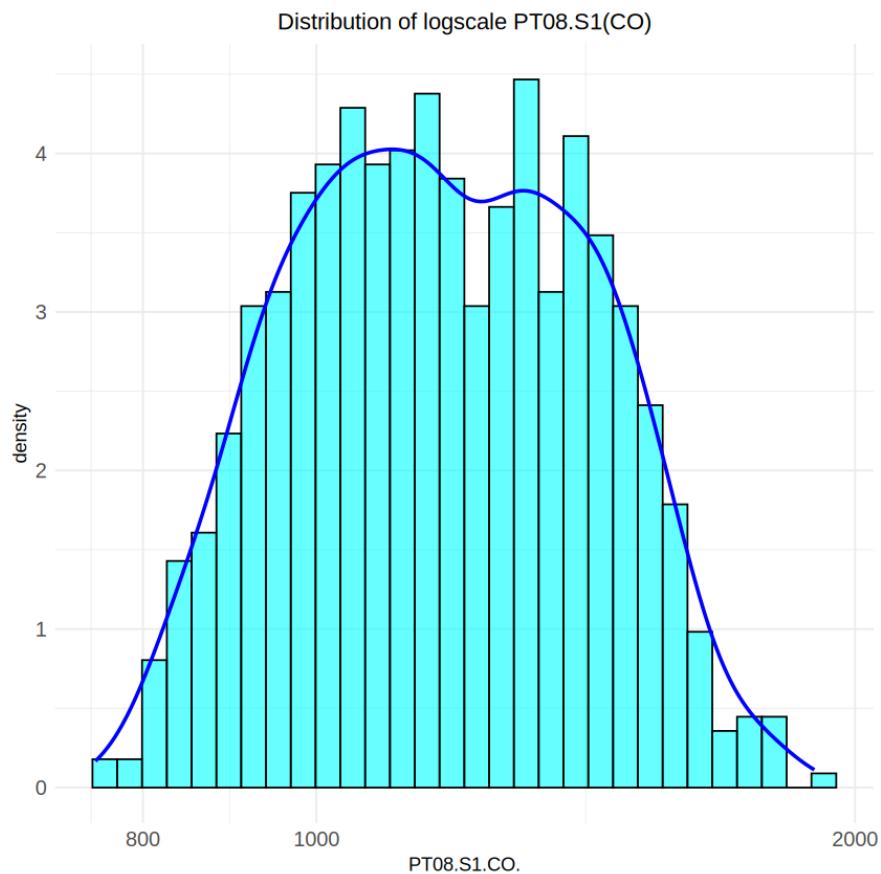


Hình 3.60: Phân phối ban đầu của Sensor response Carbon monoxide.

Nhận xét:

- Nhìn vào histogram trên, ta thấy phân phối của biến này bị lệch trái (lệch dương).

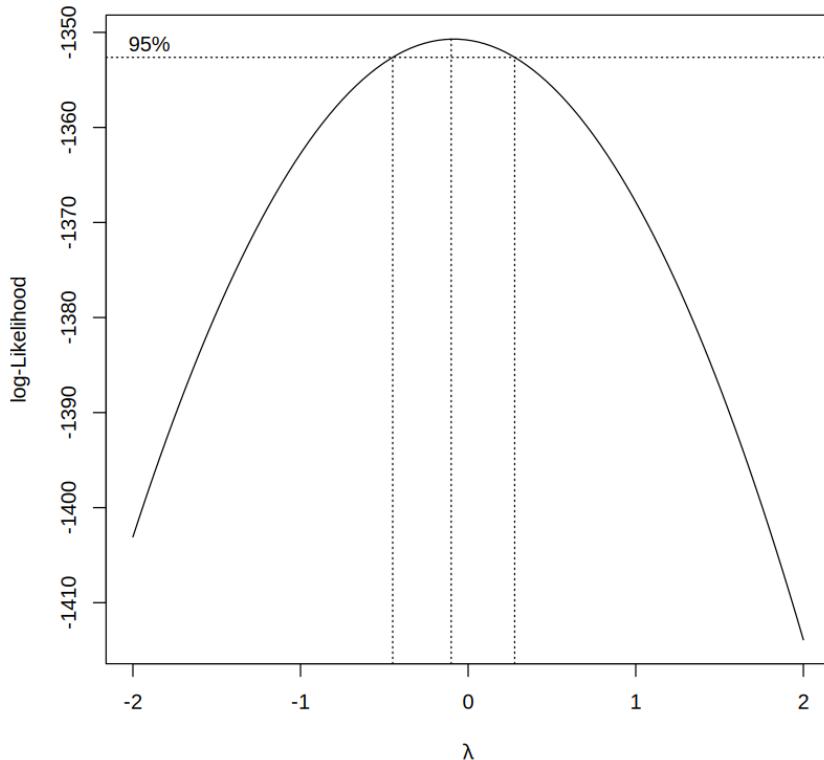
Ta thử sử dụng log-transform nó.



Hình 3.61: Phân phối sau khi log-scale của Sensor response Carbon monoxide.

Nhận xét:

- Sau khi sử dụng log-transform, hình dạng phân phối tương đối chuẩn hơn.
- Ta có thể thử sử dụng box-cox transform

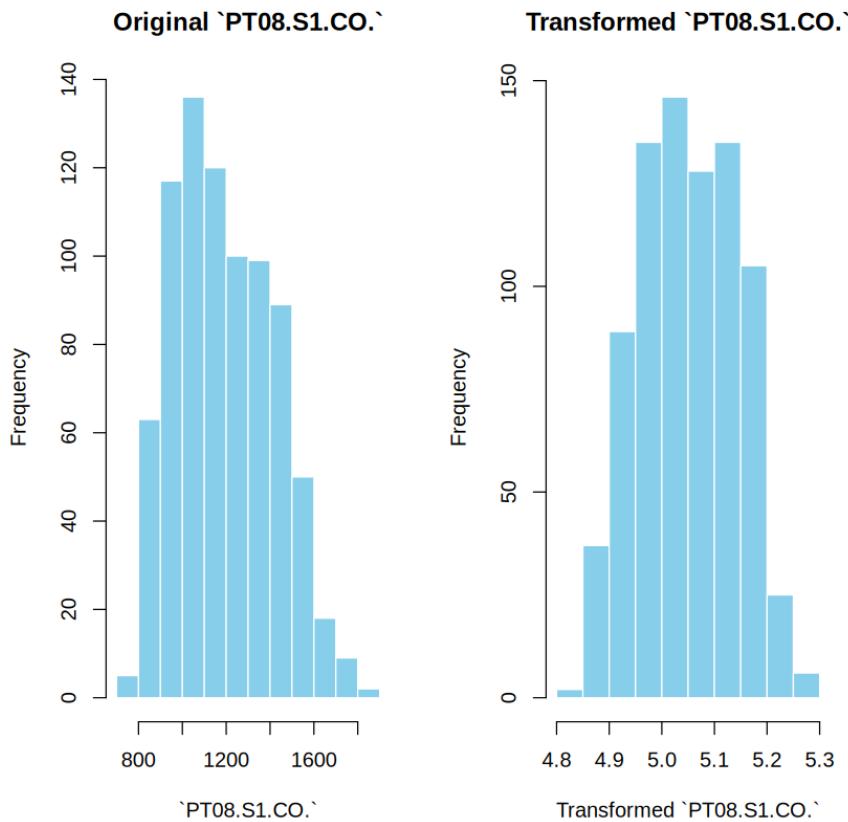


Hình 3.62: Log-likelihood với các giá trị λ của Sensor response Carbon monoxide.

Nhận xét:

- Dựa trên biểu đồ, ta tìm được giá trị lambda tối ưu với mức ý nghĩa 5% là -0.101

Và ta thực hiện biến đổi dữ liệu với giá trị lambda vừa tìm được.

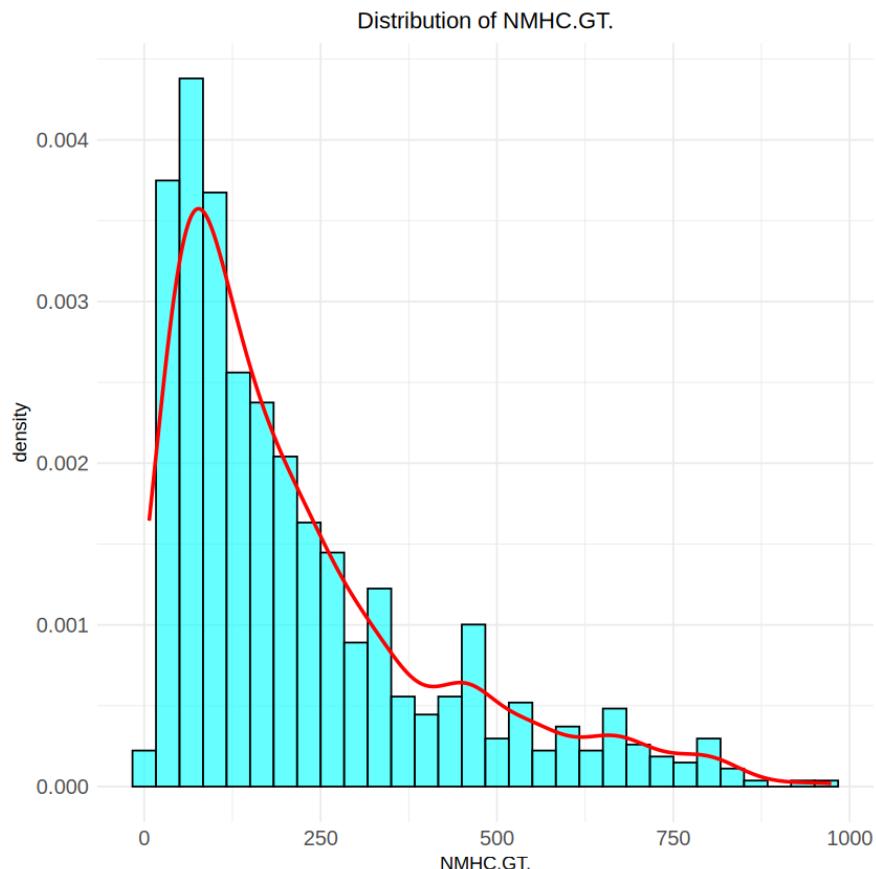


Hình 3.63: Phân phối trước và sau khi biến đổi của Sensor response Carbon monoxide.

Nhận xét:

- Ta có được giá trị lambda tối ưu là -0.101 và sử dụng giá trị này để biến đổi biến PT08.S1(CO). Biểu đồ histogram phía bên dưới thể hiện phân phối của biến này trước và sau khi biến đổi. Dễ dàng thấy được, sau khi biến đổi, biến này đã tương đối chuẩn hơn.

NMHC(GT): Non-methane hydrocarbons concentration ($\mu\text{g}/\text{m}^3$)

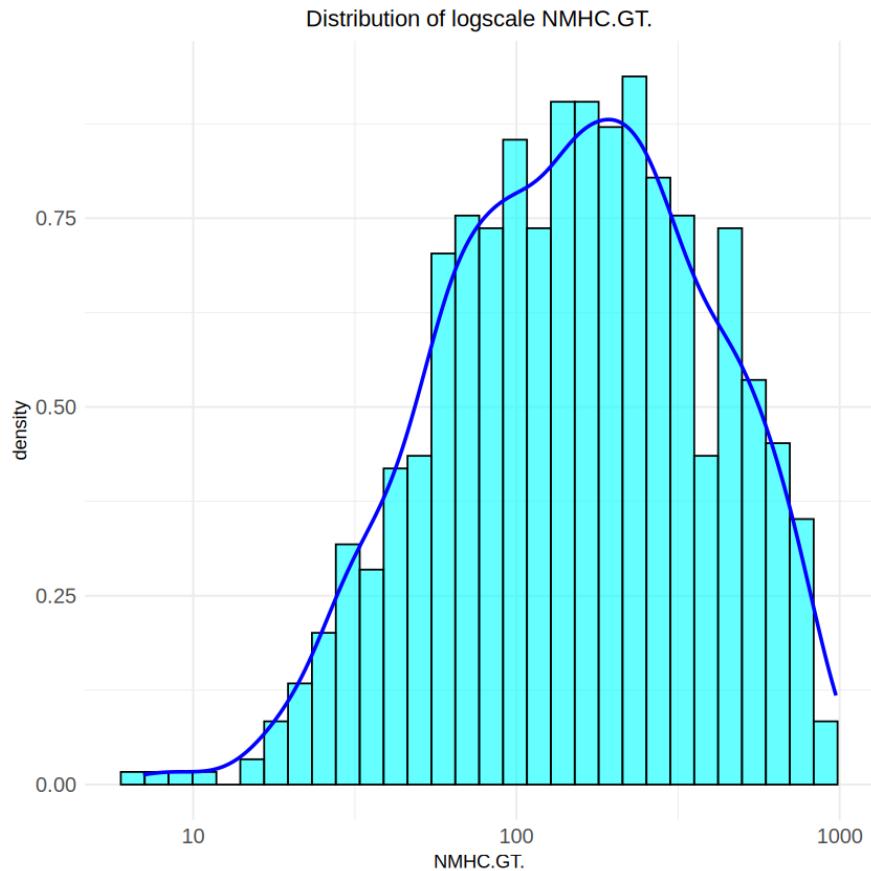


Hình 3.64: Phân phối ban đầu của Non-methane hydrocarbons.

Nhận xét:

- Phân phối tập trung ở giá trị 200

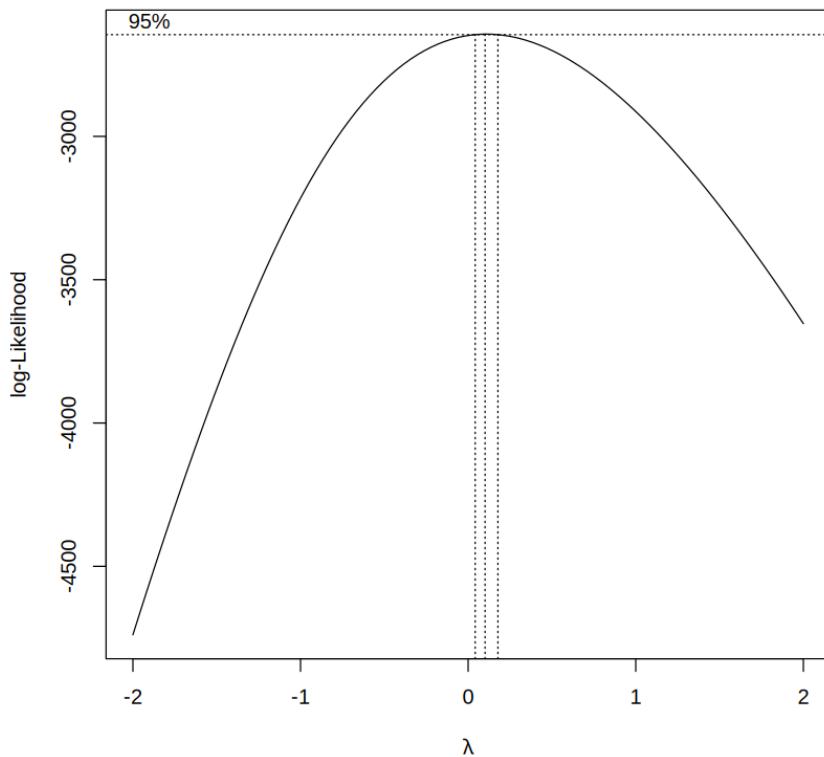
Ta thử sử dụng log-transform nó.



Hình 3.65: Phân phối sau khi log-scale của Non-methane hydrocarbons.

Nhận xét:

- Nhìn vào histogram trên, ta thấy phân phối của biến này không có nhiều thay đổi.

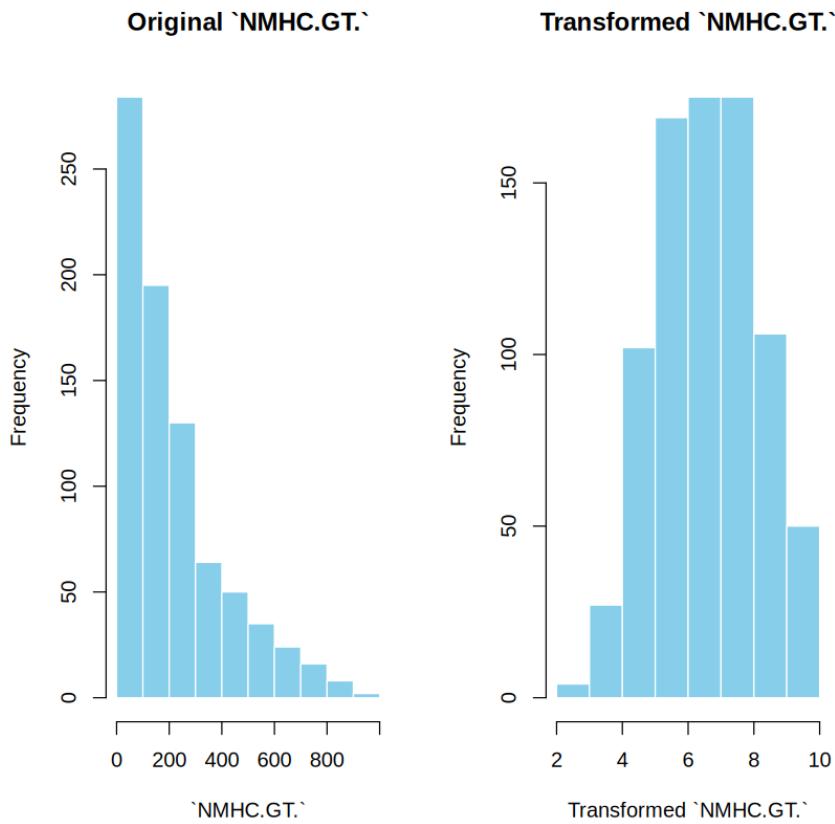


Hình 3.66: Log-likelihood với các giá trị λ của Non-methane hydrocarbons.

Nhận xét:

- Giá trị lambda phù hợp: 0.5858

Và ta thực hiện biến đổi dữ liệu với giá trị lambda vừa tìm được.

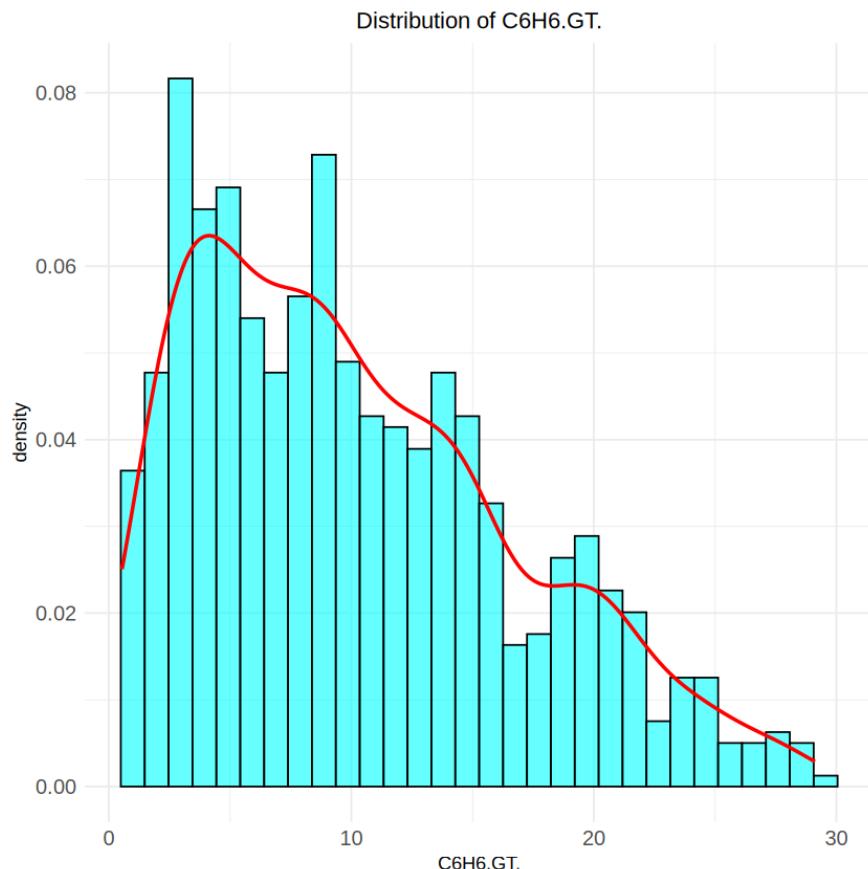


Hình 3.67: Phân phối trước và sau khi biến đổi của Non-methane hydrocarbons

Nhận xét:

- Sau khi biến đổi, giá trị NMHC(GT) tập trung nhiều ở giá trị 40

C6H6(GT): Benzene concentration ($\mu\text{g}/\text{m}^3$)

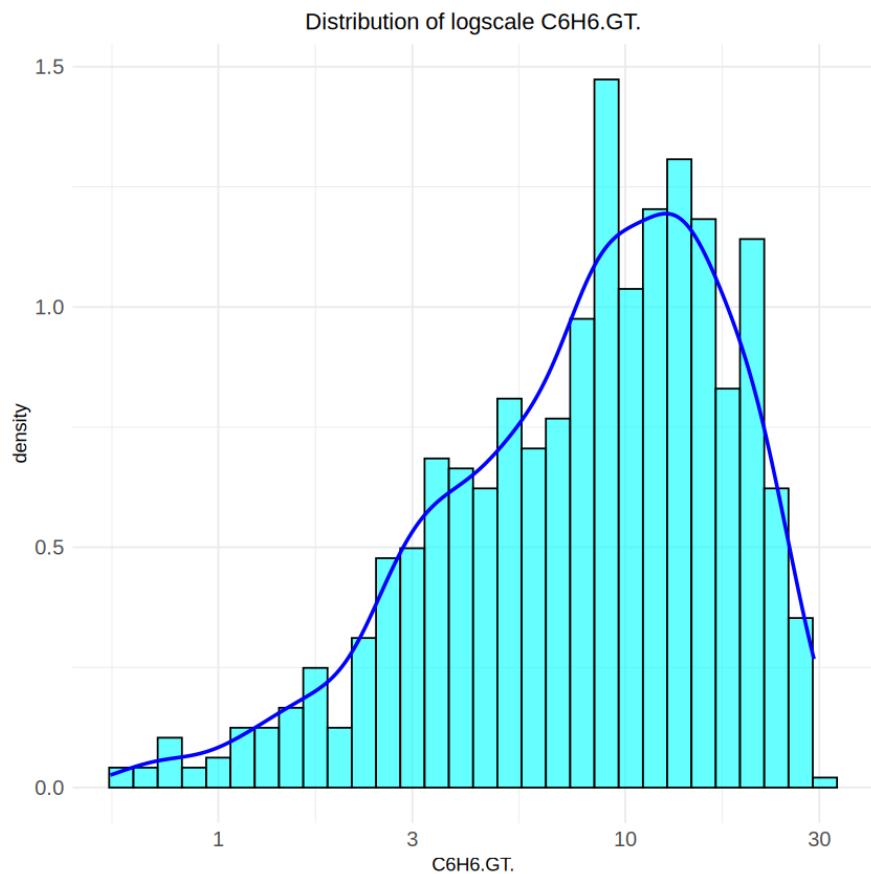


Hình 3.68: Phân phối ban đầu của Benzene concentration.

Nhận xét:

- Nhìn vào histogram trên, ta thấy phân phối của biến này bị lệch trái (lệch dương).

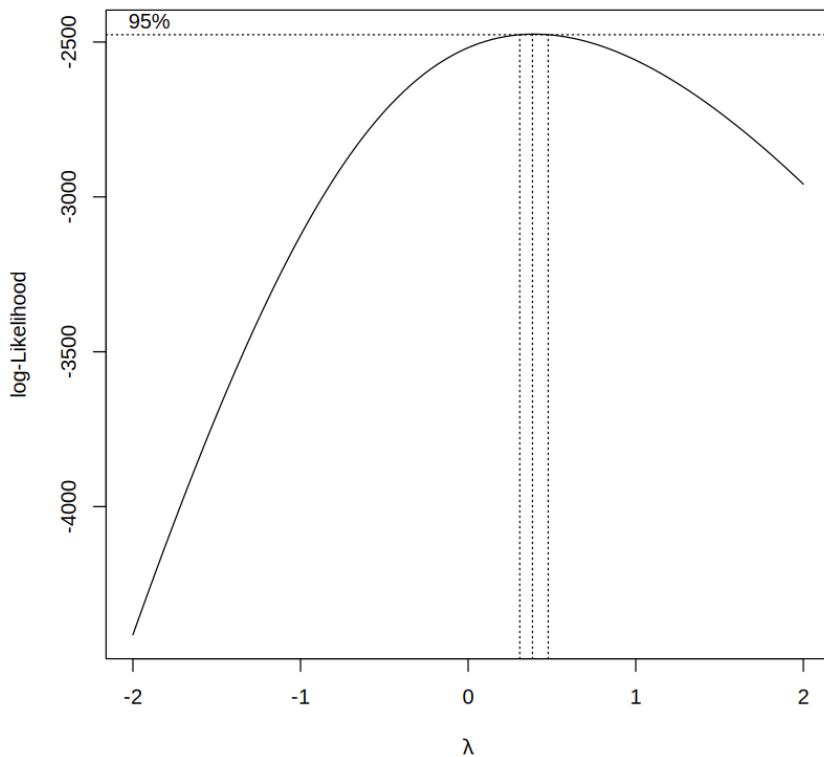
Ta thử sử dụng log-transform nó.



Hình 3.69: Phân phối sau khi log-scale của Non-methane hydrocarbons.

Nhận xét:

- Sau khi sử dụng log-transform, hình dạng phân phối tương đối chuẩn hơn.
- Ta có thể thử sử dụng box-cox transform

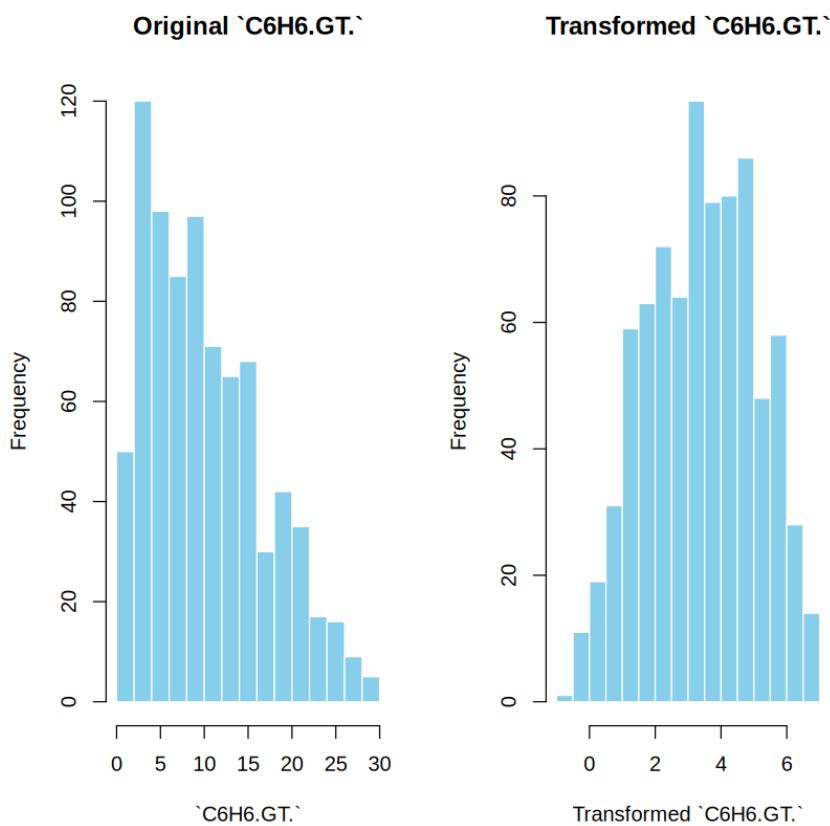


Hình 3.70: Log-likelihood với các giá trị λ của Non-methane hydrocarbons.

Nhận xét:

- Giá trị lambda phù hợp: 0.303

Và ta thực hiện biến đổi dữ liệu với giá trị lambda vừa tìm được.

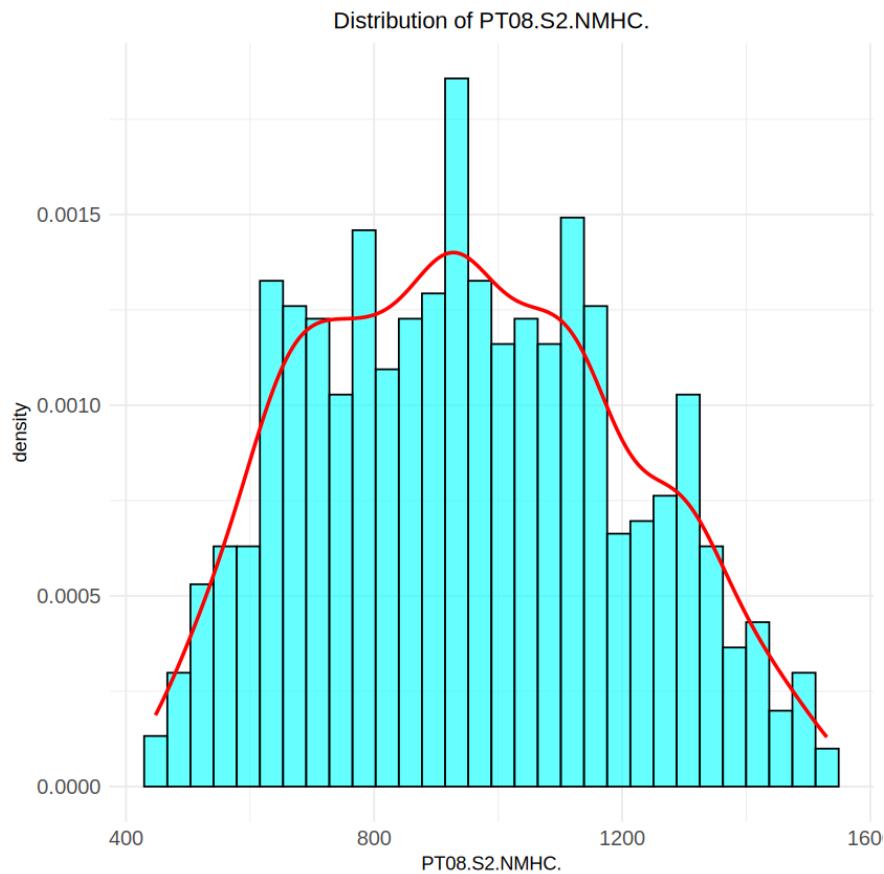


Hình 3.71: Phân phối trước và sau khi biến đổi của Non-methane hydrocarbons

Nhận xét:

- Ta có được giá trị lambda tối ưu là 0.303 và sử dụng giá trị này để biến đổi biến C6H6(GT). Biểu đồ histogram phía bên dưới thể hiện phân phối của biến này trước và sau khi biến đổi. Dễ dàng thấy được, sau khi biến đổi, biến này đã tương đối chuẩn hơn.

PT08.S2(NMHC): Sensor response for NMHC

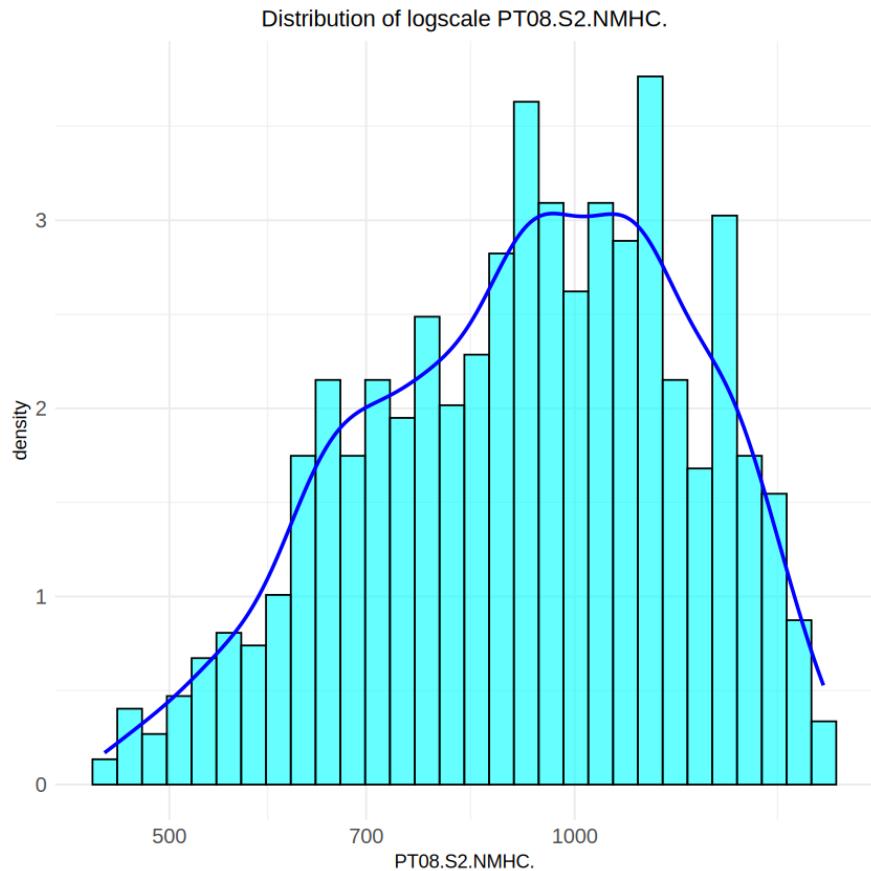


Hình 3.72: Phân phối ban đầu của Sensor response cho NMHC.

Nhận xét:

- Nhìn vào histogram trên, ta thấy phân phối của biến này tương đối chuẩn.
- Tuy nhiên, cẩn thận hơn, ta cũng chuẩn hóa biến này.

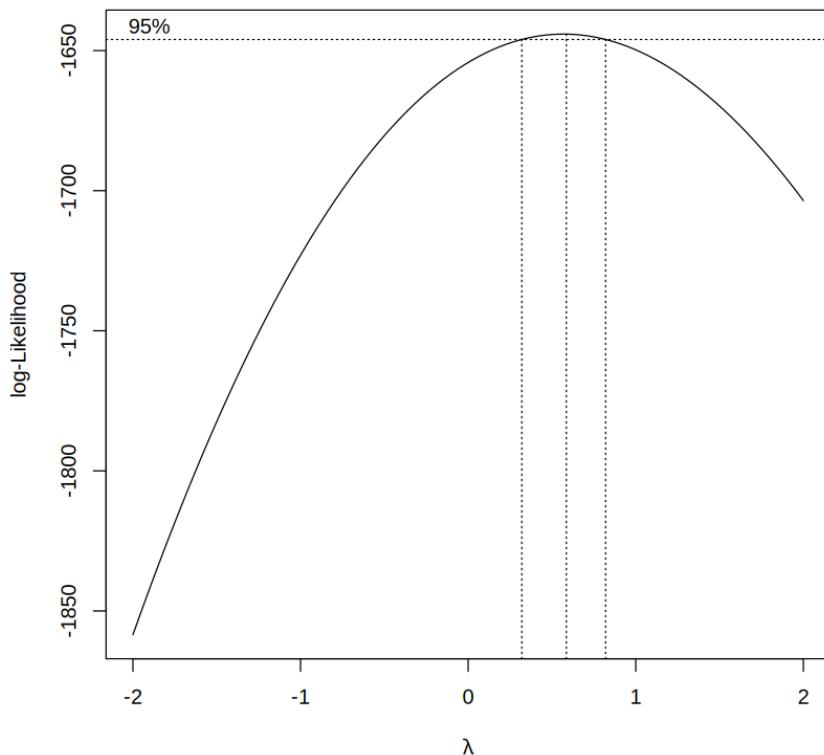
Ta thử sử dụng log-transform nó.



Hình 3.73: Phân phối sau khi log-scale của Sensor response cho NMHC.

Nhận xét:

- Sau khi log-transform, ta thấy phân phối của biến bị lệch trái.
- Ta cần sử dụng box-cox để tìm giá trị biến đổi phù hợp cho nó.

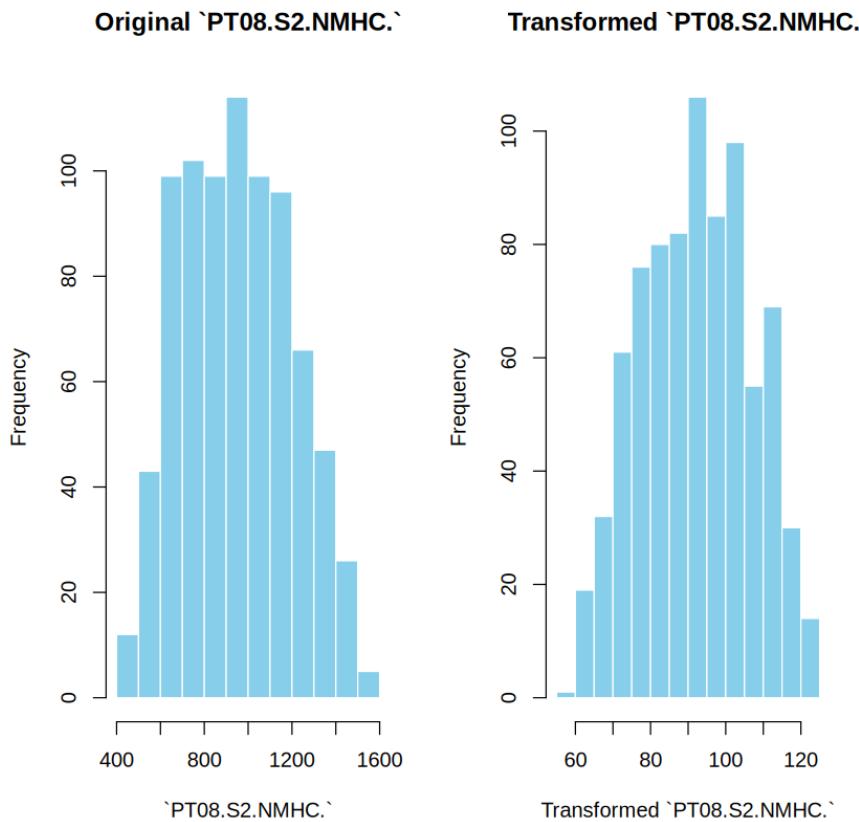


Hình 3.74: Log-likelihood với các giá trị λ của Sensor response cho NMHC.

Nhận xét:

- Giá trị lambda phù hợp với mức ý nghĩa 5% là 0.18181

Và ta thực hiện biến đổi dữ liệu với giá trị lambda vừa tìm được.

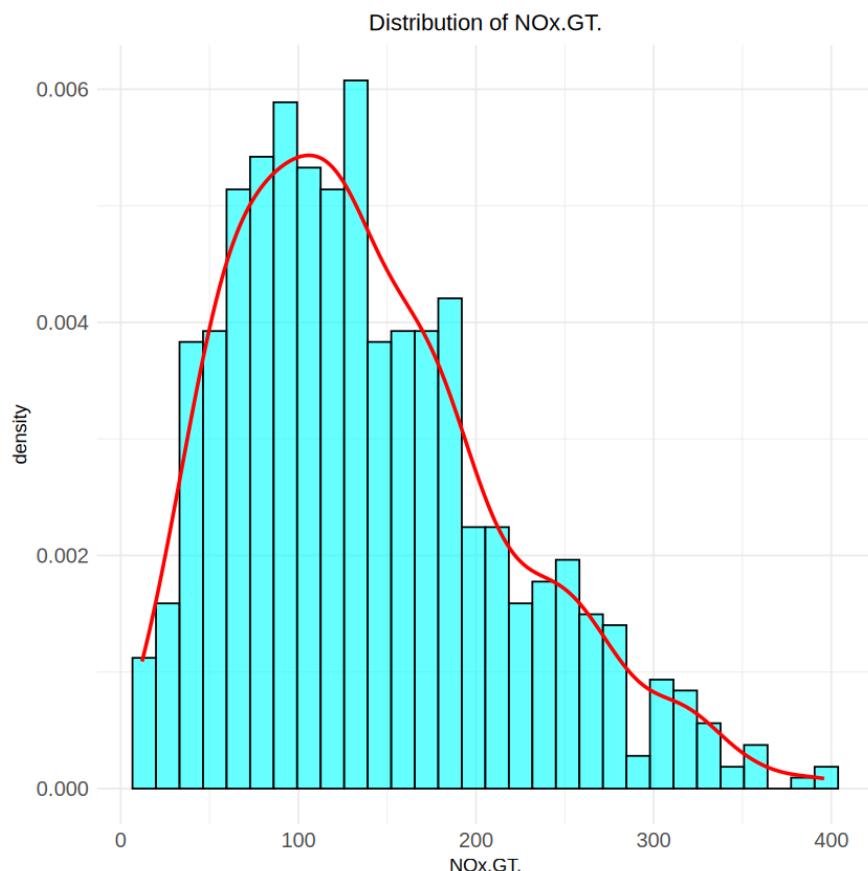


Hình 3.75: Phân phối trước và sau khi biến đổi của Sensor response cho NMHC

Nhận xét:

- Ta có được giá trị lambda tối ưu là 0.18181 và sử dụng giá trị này để biến đổi biến PT08.S2(NMHC). Biểu đồ histogram phía bên dưới thể hiện phân phối của biến này trước và sau khi biến đổi. Dễ dàng thấy được, sau khi biến đổi, biến này đã tương đối chuẩn hơn.

$NOx(GT)$: Nitrogen oxides concentration (ppb)

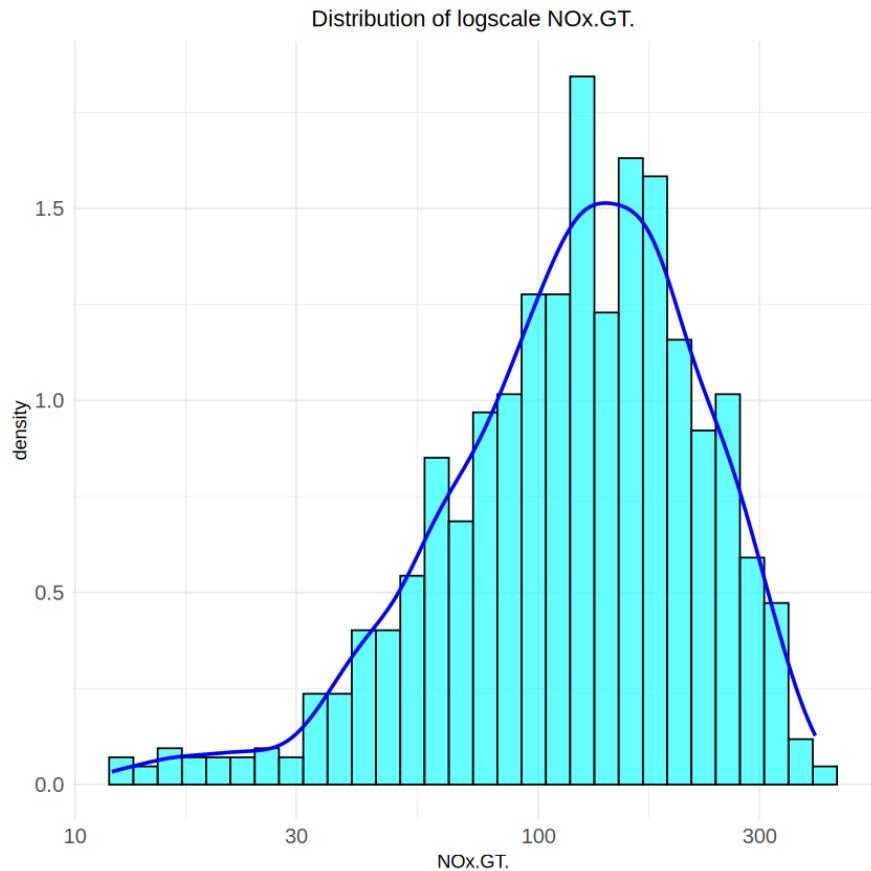


Hình 3.76: Phân phối ban đầu của Sensor response cho Nitrogen oxides.

Nhận xét:

- Nhìn vào histogram trên, ta thấy phân phối của biến này bị lệch trái (lệch dương).

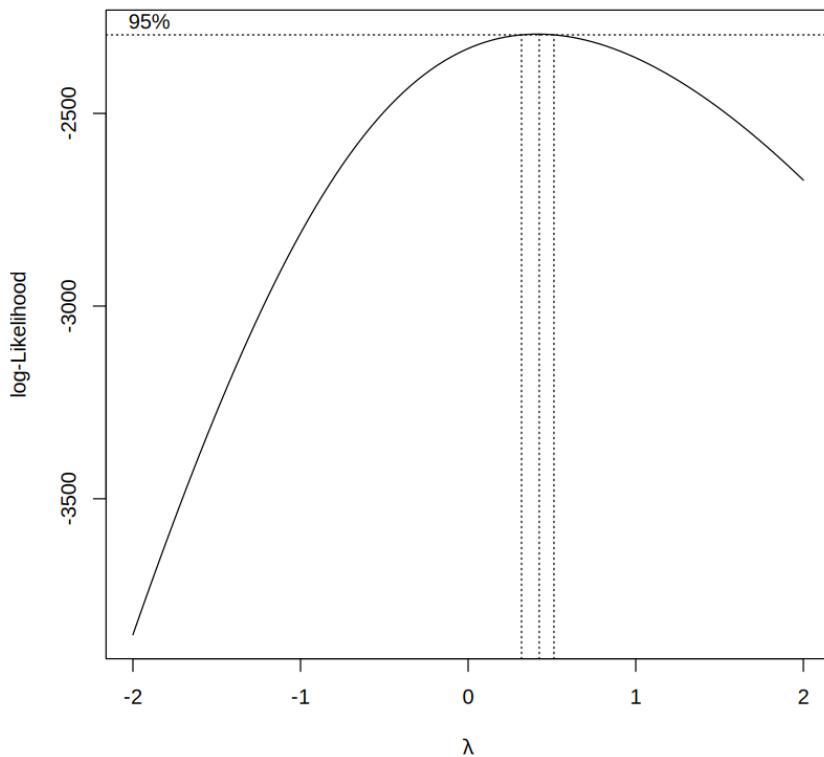
Ta thử sử dụng log-transform nó.



Hình 3.77: Phân phối sau khi log-scale của Sensor response cho Nitrogen oxides.

Nhận xét:

- Sau khi logscale, ta thấy phân phối có chiều hướng lệch phải (lệch âm)
- Ta cần sử dụng box-cox

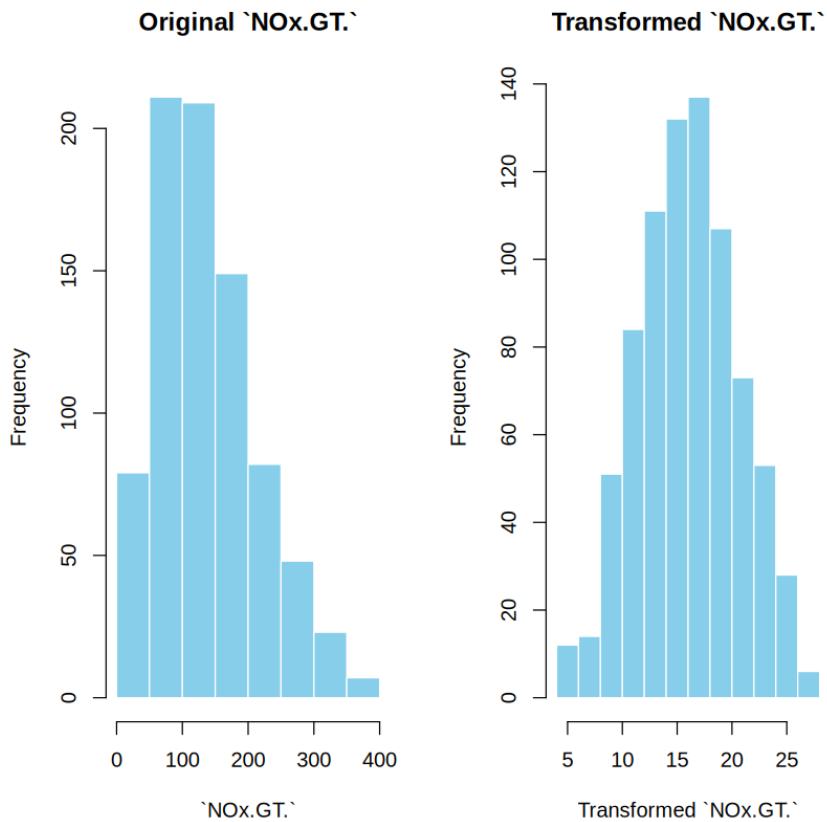


Hình 3.78: Log-likelihood với các giá trị λ của Sensor response cho Nitrogen oxides.

Nhận xét:

- Với mức ý nghĩa 5%, ta tìm được giá trị lambda phù hợp là 0.2222

Và ta thực hiện biến đổi dữ liệu với giá trị lambda vừa tìm được.

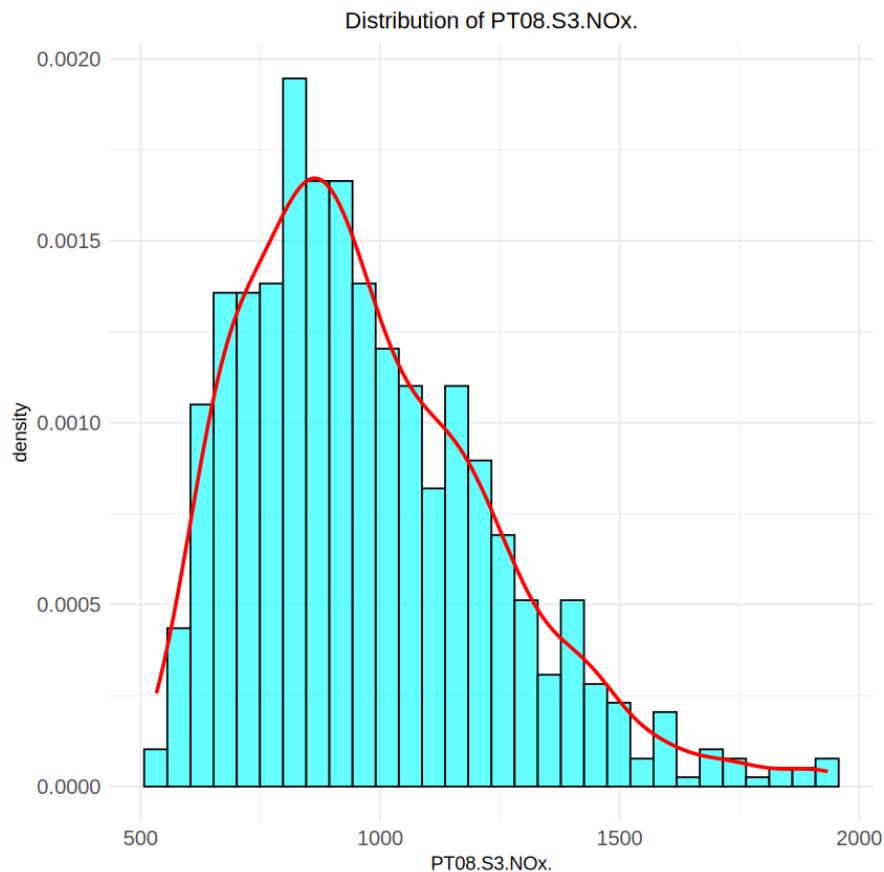


Hình 3.79: Phân phối trước và sau khi biến đổi của Sensor response cho Nitrogen oxides.

Nhận xét:

- Ta có được giá trị lambda tối ưu là 0.2222 và sử dụng giá trị này để biến đổi biến NOx(GT). Biểu đồ histogram phía bên dưới thể hiện phân phối của biến này trước và sau khi biến đổi. Dễ dàng thấy được, sau khi biến đổi, biến này đã tương đối chuẩn hơn.

PT08.S3(NO_x): Sensor response for NO_x

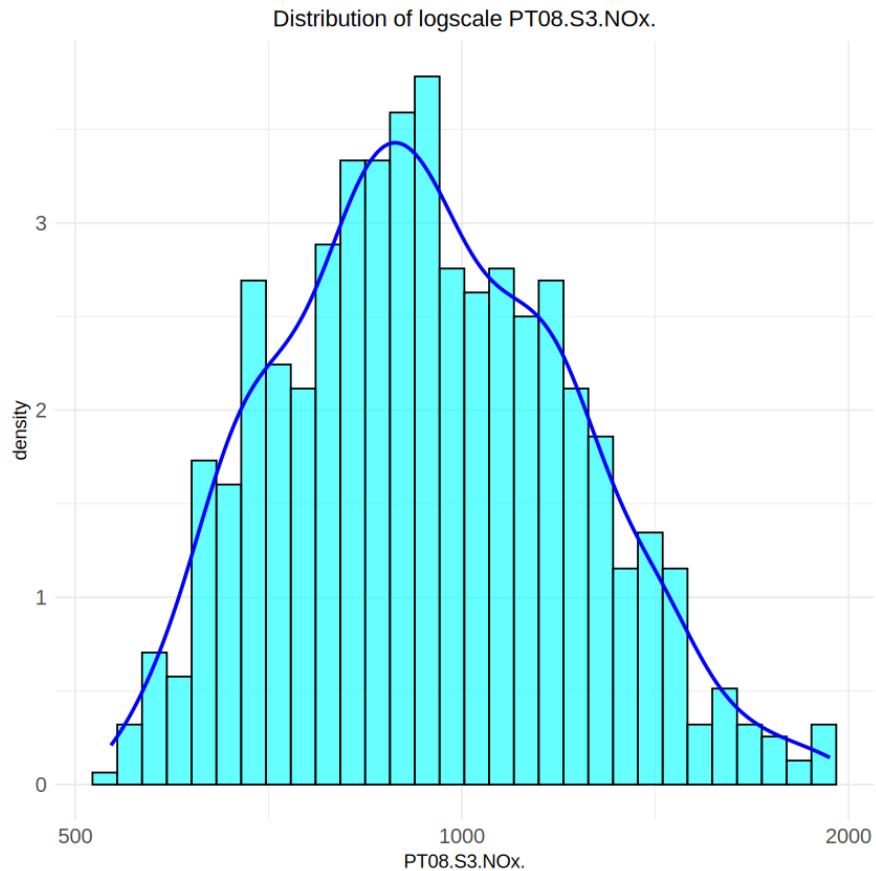


Hình 3.80: Phân phối ban đầu của Sensor response cho Nitrogen oxides.

Nhận xét:

- Nhìn vào histogram trên, ta thấy phân phối của biến này bị lệch trái (lệch dương).

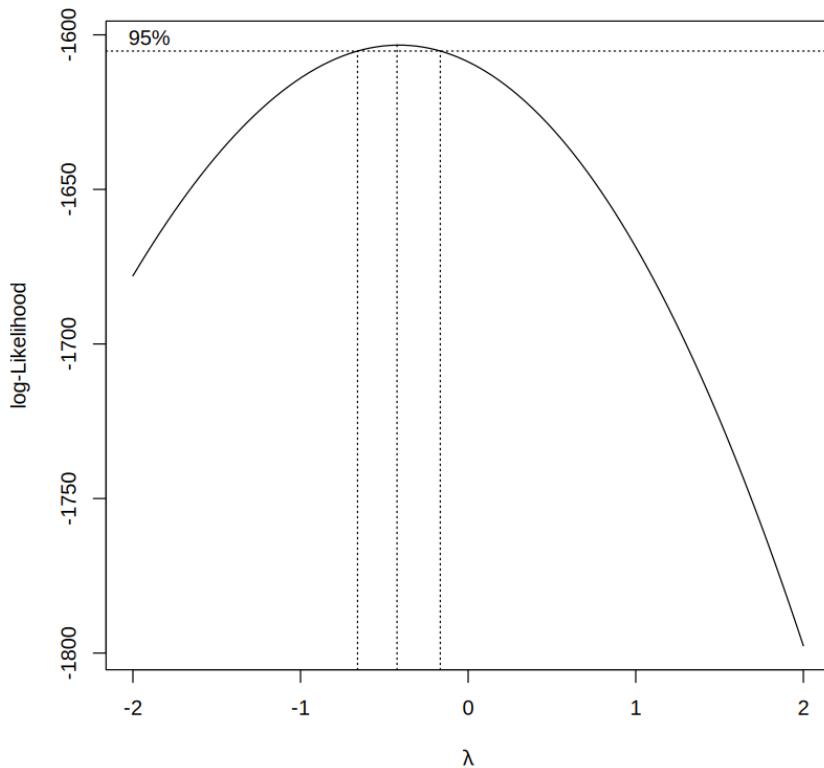
Ta thử sử dụng log-transform nó.



Hình 3.81: Phân phối sau khi log-scale của Sensor response cho Nitrogen oxides.

Nhận xét:

- Sau khi logscale, ta thấy phân phối của biến này xấp xỉ chuẩn hơn

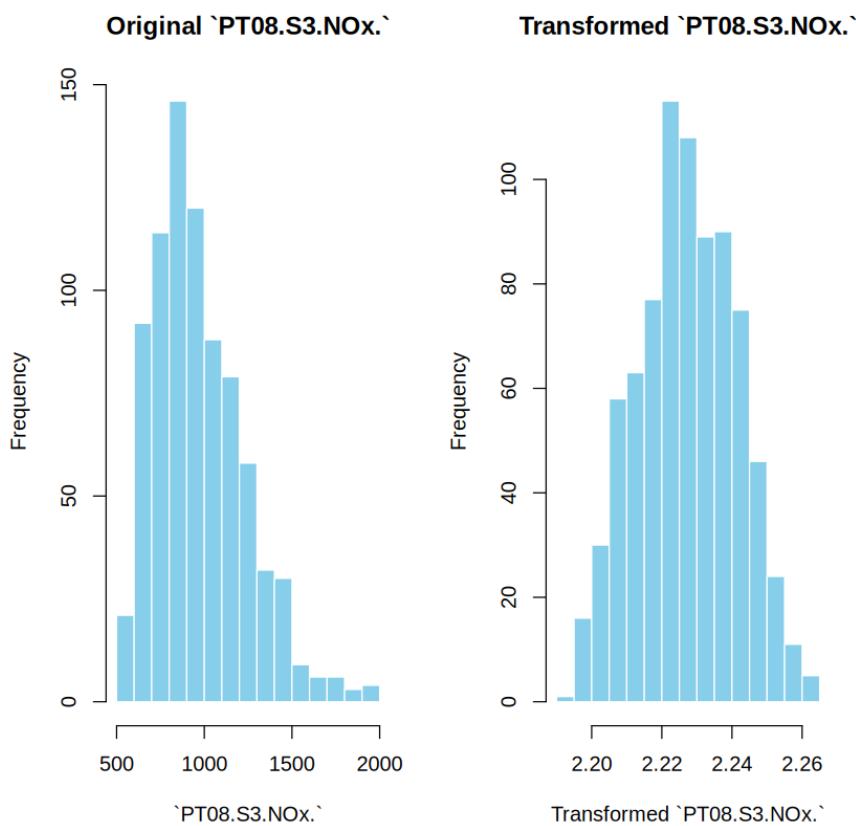


Hình 3.82: Log-likelihood với các giá trị λ của Sensor response cho Nitrogen oxides.

Nhận xét:

- Với mức ý nghĩa 5%, ta tìm được giá trị lambda phù hợp là -0.02020

Và ta thực hiện biến đổi dữ liệu với giá trị lambda vừa tìm được.

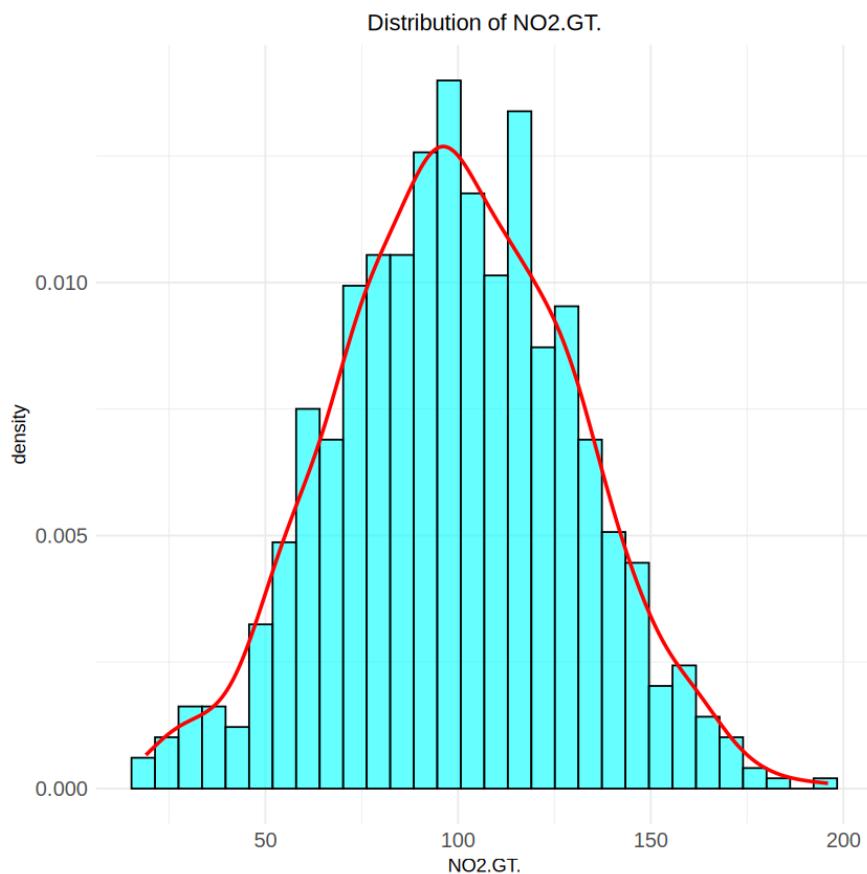


Hình 3.83: Phân phối trước và sau khi biến đổi của Sensor response cho Nitrogen oxides.

Nhận xét:

- Ta có được giá trị lambda tối ưu là -0.02020 và sử dụng giá trị này để biến đổi biến PT08.S3(NOx). Biểu đồ histogram phía bên dưới thể hiện phân phối của biến này trước và sau khi biến đổi. Dễ dàng thấy được, sau khi biến đổi, biến này đã tương đối chuẩn hơn.

NO₂(GT): Nitrogen dioxide concentration ($\mu\text{g}/\text{m}^3$)

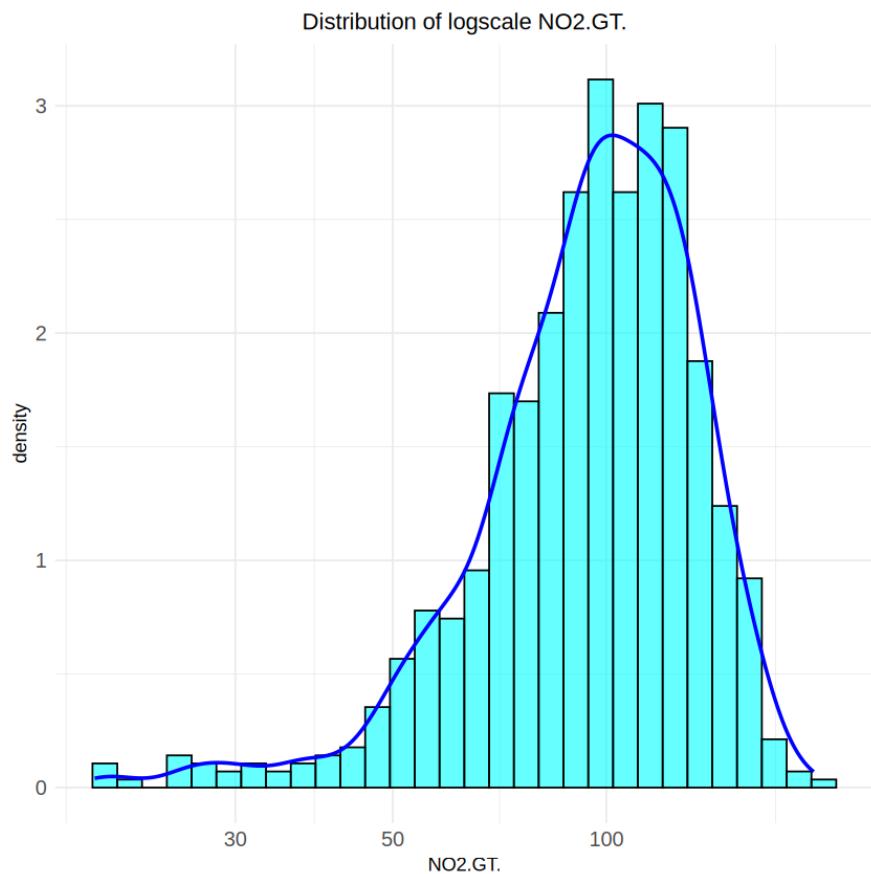


Hình 3.84: Phân phối ban đầu của Sensor response cho Nitrogen oxides.

Nhận xét:

- Nhìn vào histogram trên, ta thấy phân phối của biến này xấp xỉ chuẩn.

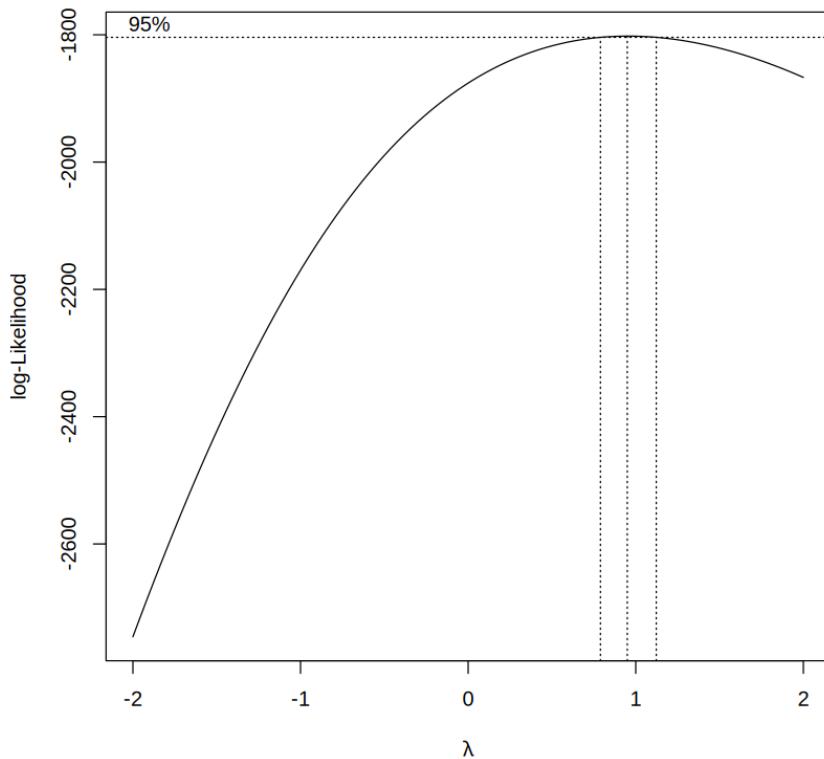
Ta thử sử dụng log-transform nó.



Hình 3.85: Phân phối sau khi log-scale của Sensor response cho Nitrogen oxides.

Nhận xét:

- Sau khi logscale, ta thấy phân phối của biến này có chiều hướng lệch phải (lệch âm)

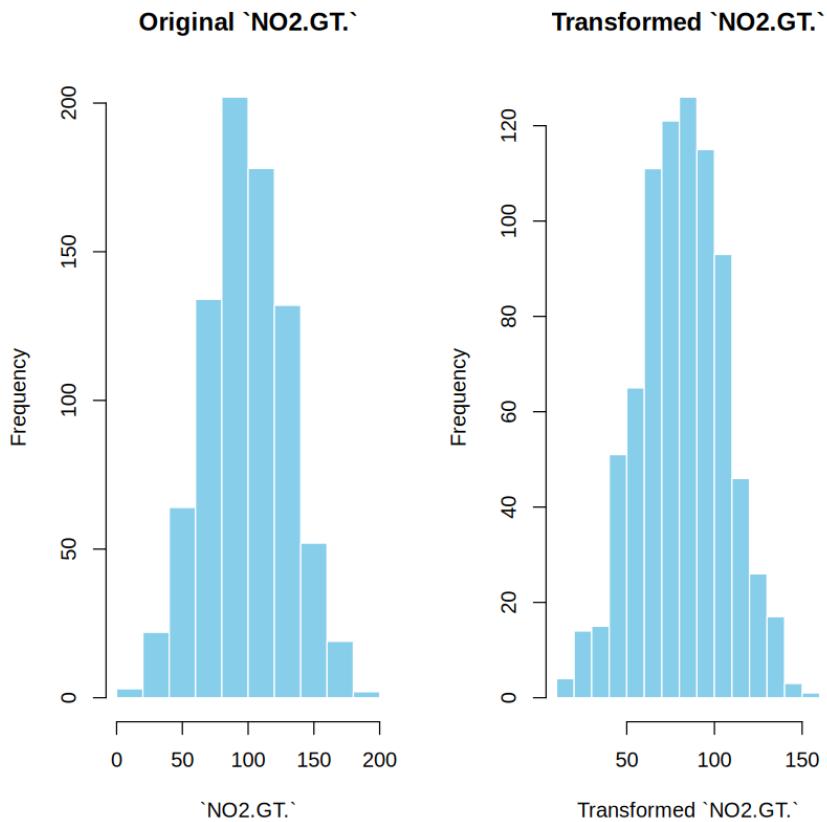


Hình 3.86: Log-likelihood với các giá trị λ của Sensor response cho Nitrogen oxides.

Nhận xét:

- Với mức ý nghĩa 5%, ta tìm được giá trị lambda phù hợp 0.58585

Và ta thực hiện biến đổi dữ liệu với giá trị lambda vừa tìm được.

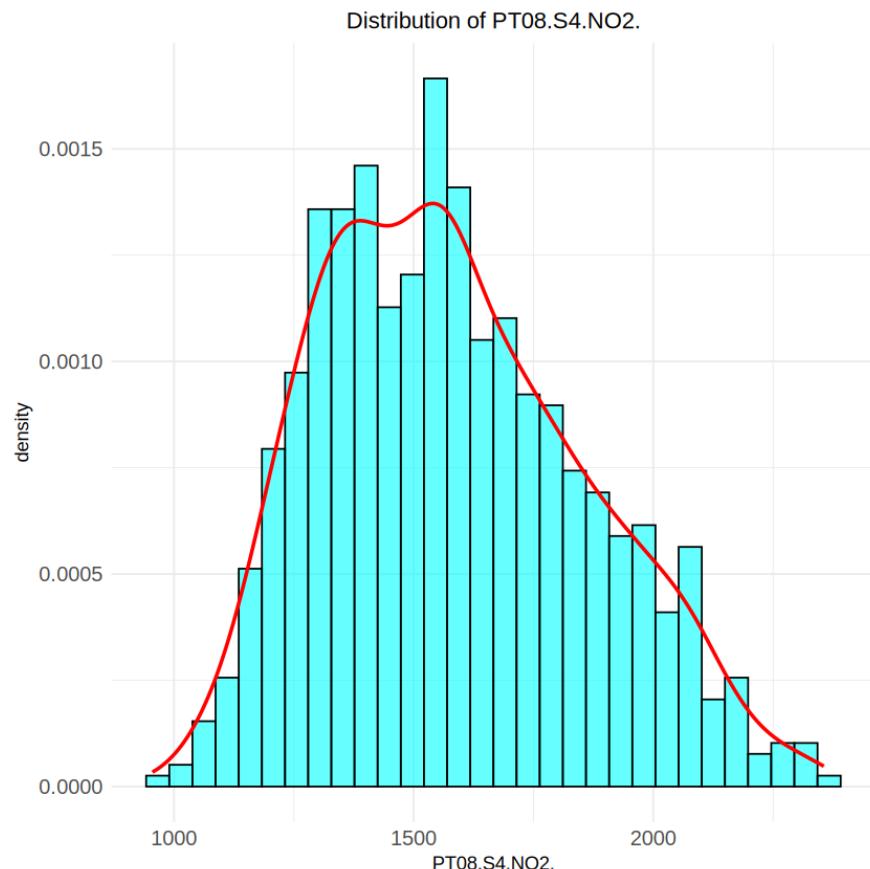


Hình 3.87: Phân phối trước và sau khi biến đổi của Sensor response cho Nitrogen oxides.

Nhận xét:

- Ta có được giá trị lambda tối ưu là 0.58585 và sử dụng giá trị này để biến đổi biến NO2(GT). Biểu đồ histogram phía bên dưới thể hiện phân phối của biến này trước và sau khi biến đổi. Dễ dàng thấy được, sau khi biến đổi, biến này đã tương đối chuẩn hơn.

PT08.S4(NO₂): Sensor response for NO₂

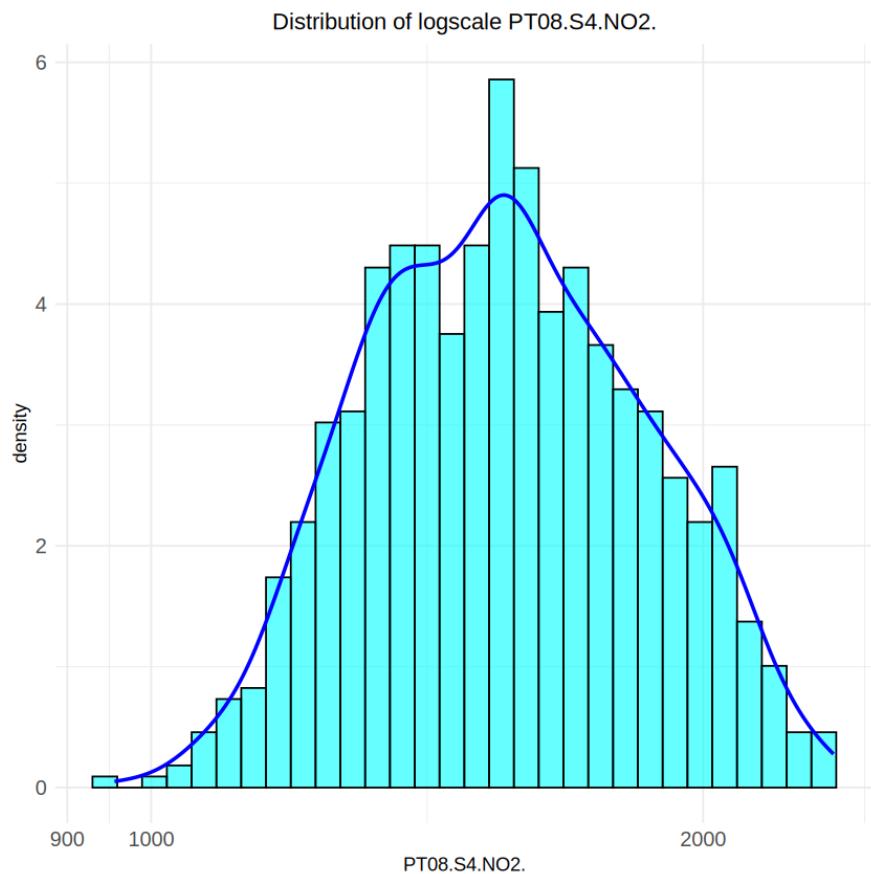


Hình 3.88: Phân phối ban đầu của Sensor response cho Nitrogen oxides.

Nhận xét:

- Nhìn vào histogram trên, ta thấy phân phối của biến này xấp xỉ chuẩn.

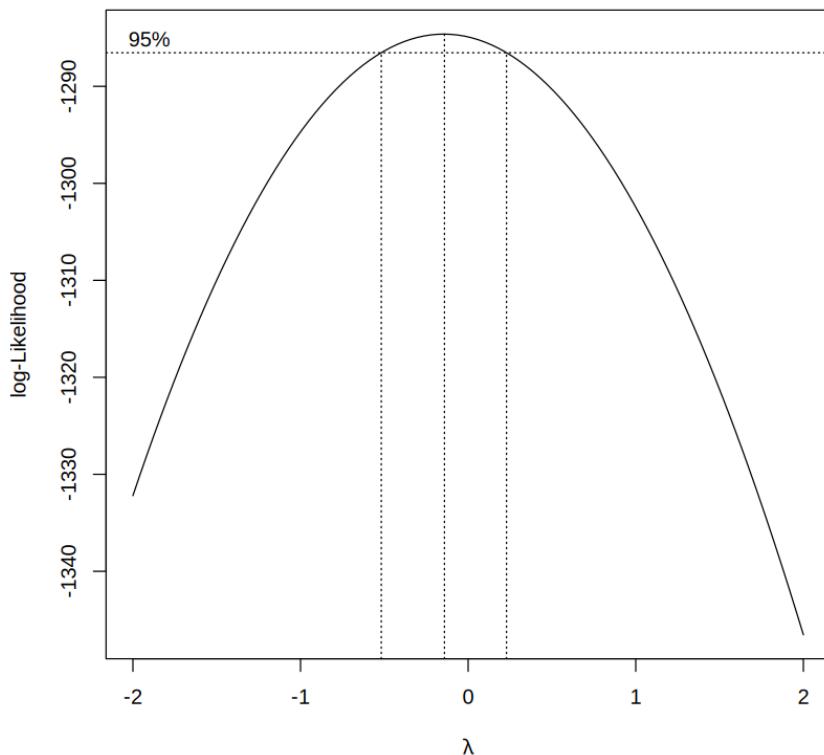
Ta thử sử dụng log-transform nó.



Hình 3.89: Phân phối sau khi log-scale của Sensor response cho Nitrogen oxides.

Nhận xét:

- Sau khi logscale, ta vẫn thấy phân phối của biến này vẫn xấp xỉ chuẩn

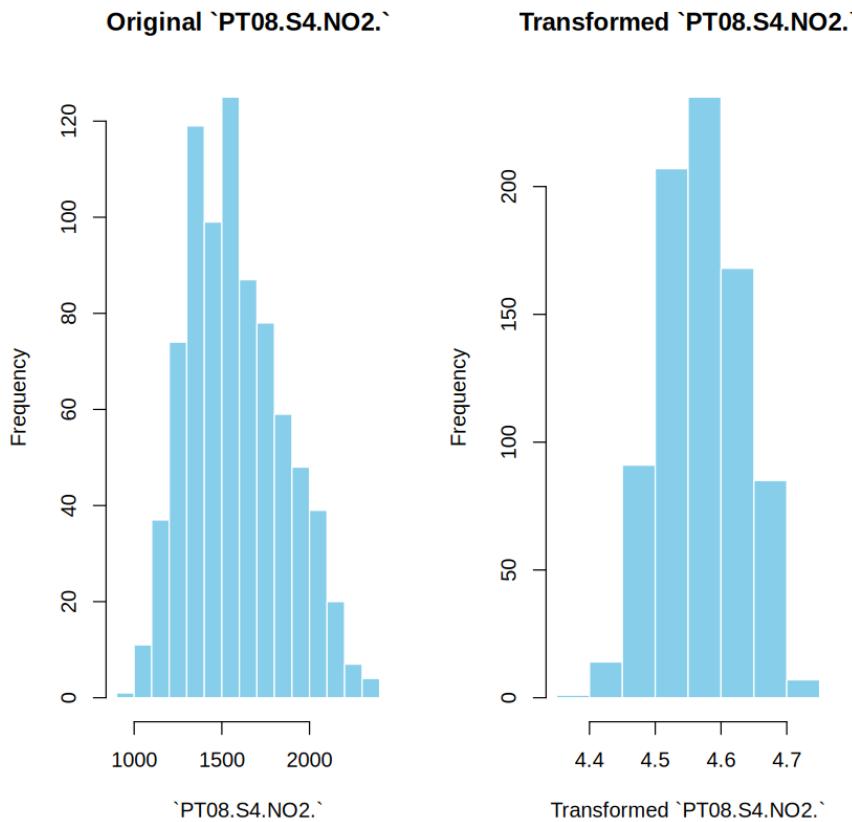


Hình 3.90: Log-likelihood với các giá trị λ của Sensor response cho Nitrogen oxides.

Nhận xét:

- Với mức ý nghĩa 5%, ta tìm được giá trị lambda phù hợp là 0.7070707

Và ta thực hiện biến đổi dữ liệu với giá trị lambda vừa tìm được.

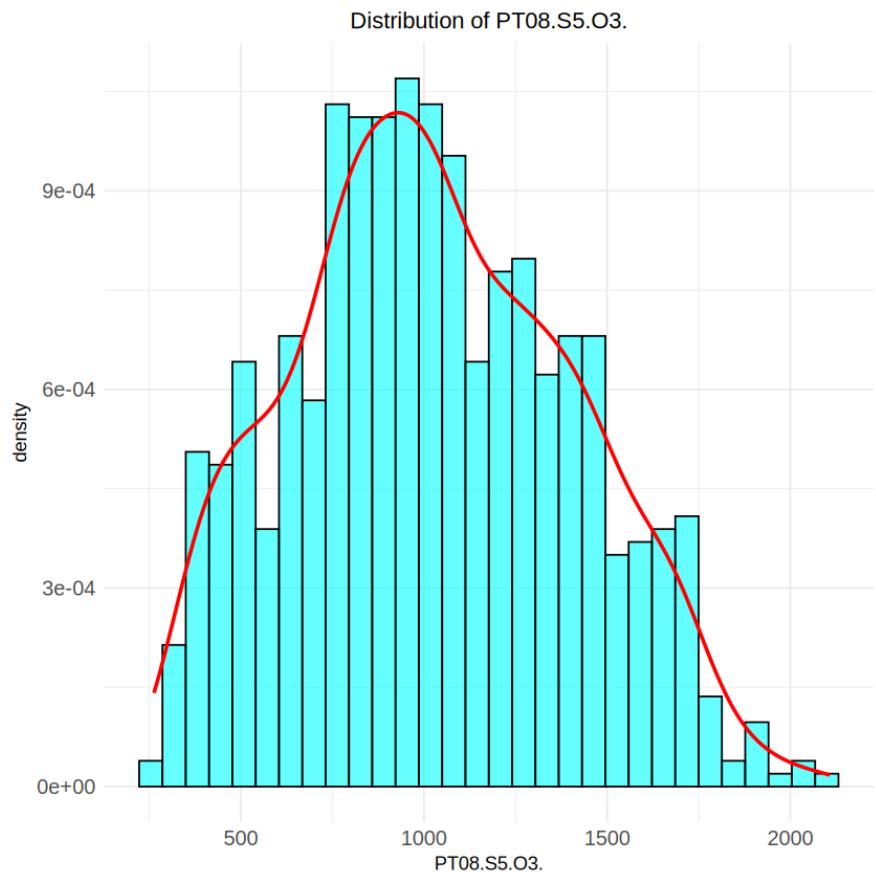


Hình 3.91: Phân phối trước và sau khi biến đổi của Sensor response cho Nitrogen oxides.

Nhận xét:

- Ta có được giá trị lambda tối ưu là 0.7070707 và sử dụng giá trị này để biến đổi biến PT08.S4(NO2). Biểu đồ histogram phía bên dưới thể hiện phân phối của biến này trước và sau khi biến đổi. Dễ dàng thấy được, sau khi biến đổi, biến này đã tương đối chuẩn hơn.

PT08.S5(O3): Sensor response for ozone

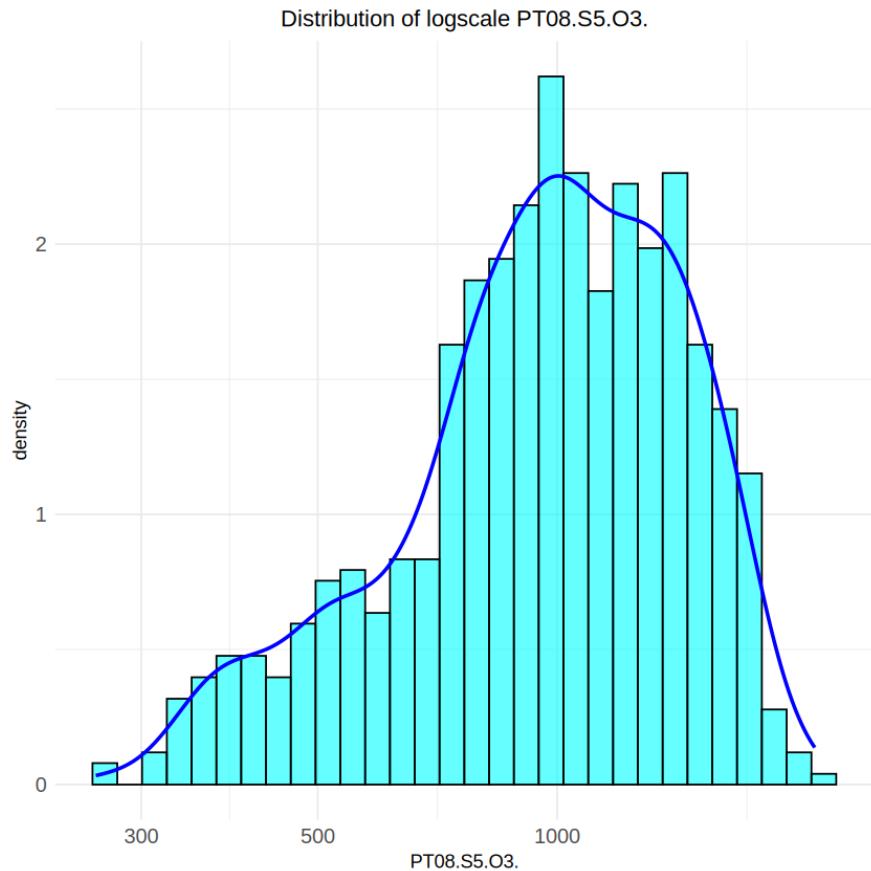


Hình 3.92: Phân phối ban đầu của Sensor response cho Nitrogen oxides.

Nhận xét:

- Nhìn vào histogram trên, ta thấy phân phối của biến này bị lệch trái (lệch dương).

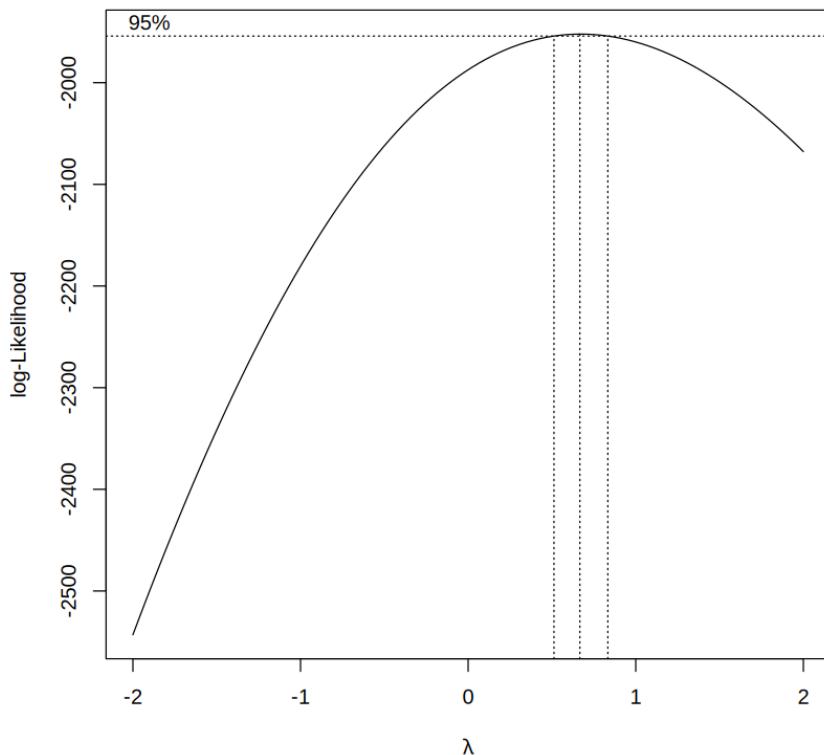
Ta thử sử dụng log-transform nó.



Hình 3.93: Phân phối sau khi log-scale của Sensor response cho Nitrogen oxides.

Nhận xét:

- Sau khi logscale, ta thấy phân phối của biến này có chiều hướng lệch phải (lệch âm)

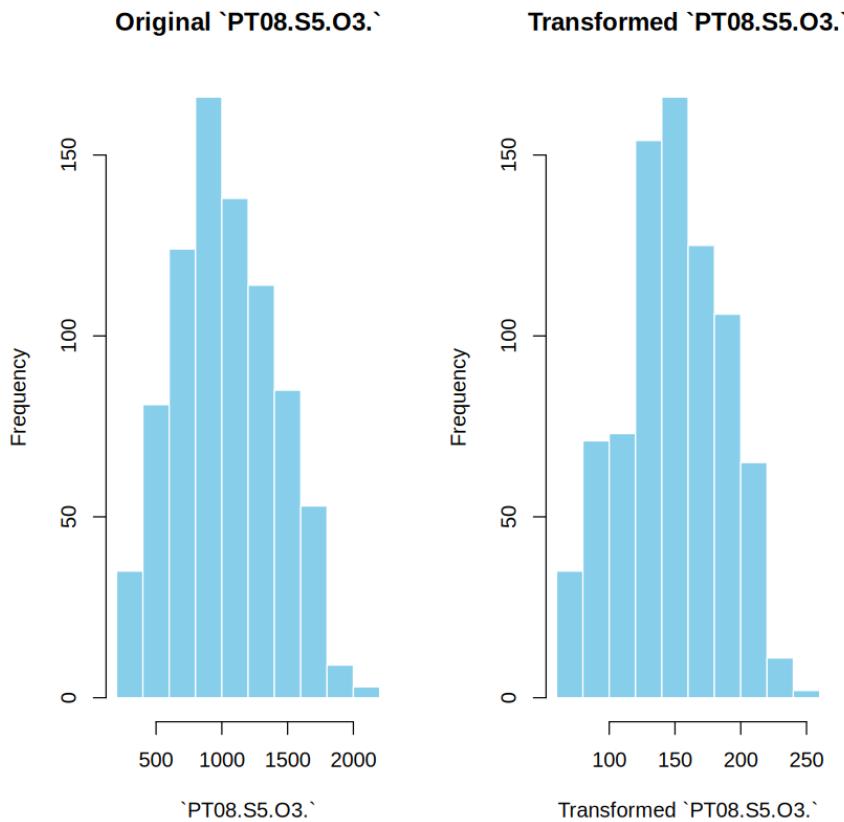


Hình 3.94: Log-likelihood với các giá trị λ của Sensor response cho Nitrogen oxides.

Nhận xét:

- Với mức ý nghĩa 5%, ta tìm được giá trị lambda phù hợp 0.3434

Và ta thực hiện biến đổi dữ liệu với giá trị lambda vừa tìm được.

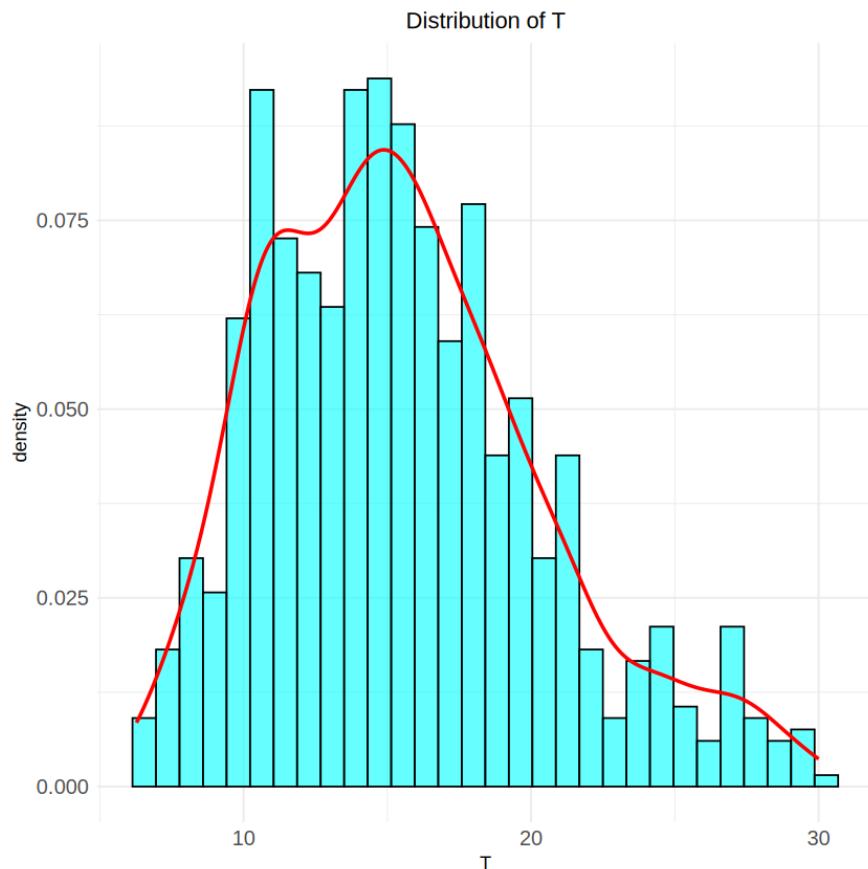


Hình 3.95: Phân phối trước và sau khi biến đổi của Sensor response cho Nitrogen oxides.

Nhận xét:

- Ta có được giá trị lambda tối ưu là 0.3434 và sử dụng giá trị này để biến đổi biến PT08.S5(O3). Biểu đồ histogram phía bên dưới thể hiện phân phối của biến này trước và sau khi biến đổi. Dễ dàng thấy được, sau khi biến đổi, biến này đã tương đối chuẩn hơn.

T : Temperature (Celsius)

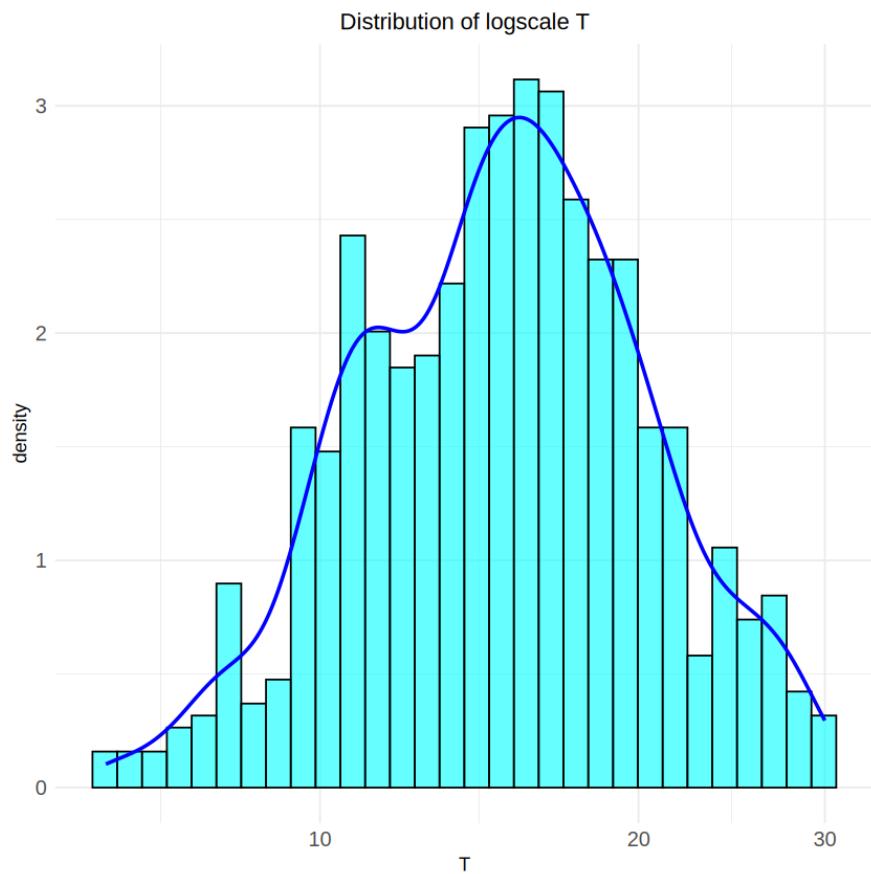


Hình 3.96: Phân phối ban đầu của Sensor response cho Nitrogen oxides.

Nhận xét:

- Nhìn vào histogram trên, ta thấy phân phối của biến này bị lệch trái (lệch âm).

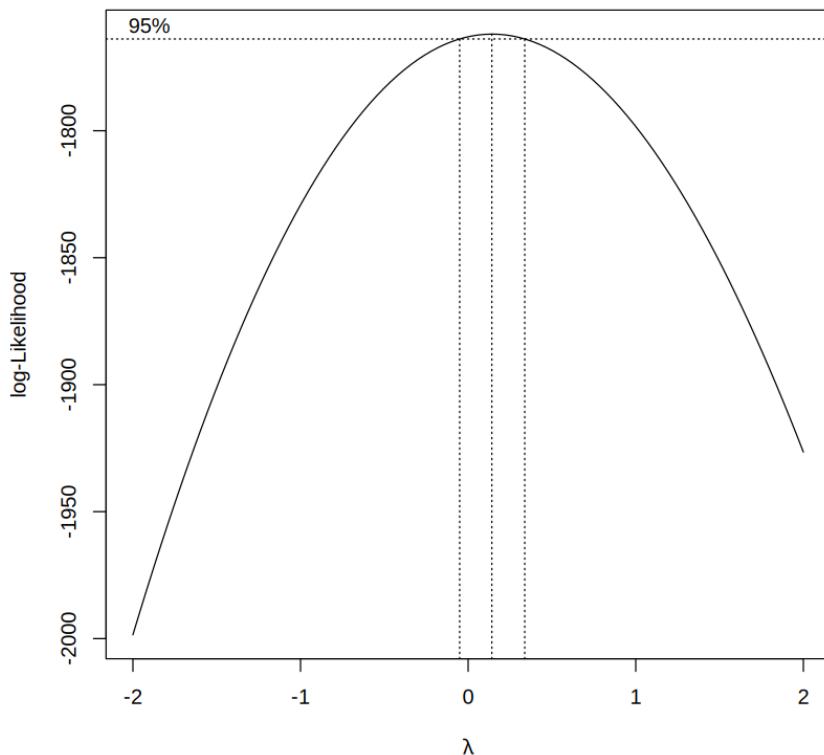
Ta thử sử dụng log-transform nó.



Hình 3.97: Phân phối sau khi log-scale của Sensor response cho Nitrogen oxides.

Nhận xét:

- Sau khi logscale, ta thấy phân phối của biến này có chiều hướng lệch phải (lệch âm)

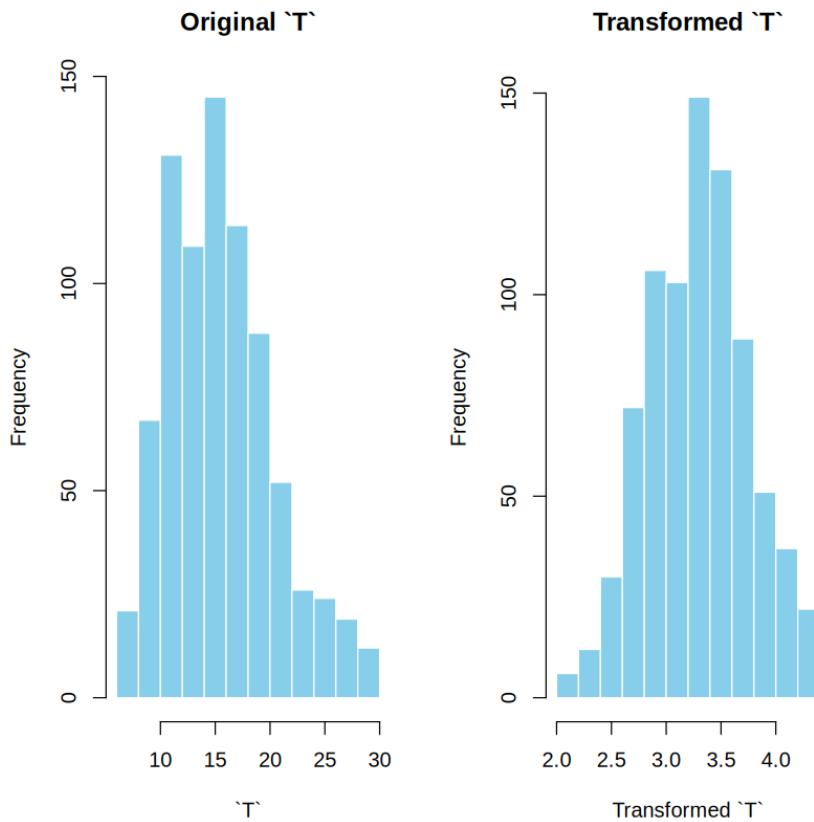


Hình 3.98: Log-likelihood với các giá trị λ của Sensor response cho Nitrogen oxides.

Nhận xét:

- Với mức ý nghĩa 5%, ta tìm được giá trị lambda phù hợp là 0.62626

Và ta thực hiện biến đổi dữ liệu với giá trị lambda vừa tìm được.

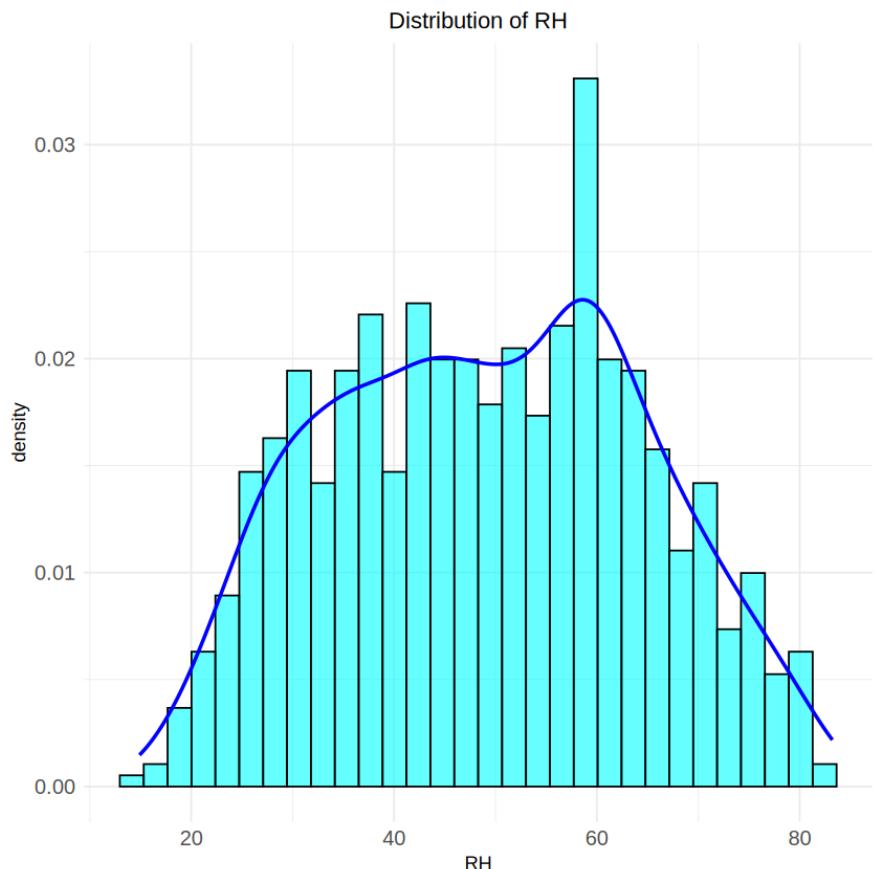


Hình 3.99: Phân phối trước và sau khi biến đổi của Sensor response cho Nitrogen oxides.

Nhận xét:

- Ta có được giá trị lambda tối ưu là 0.62626 và sử dụng giá trị này để biến đổi biến T. Biểu đồ histogram phía bên dưới thể hiện phân phối của biến này trước và sau khi biến đổi. Dễ dàng thấy được, sau khi biến đổi, biến này đã tương đối chuẩn hơn.

RH: Relative humidity (%)

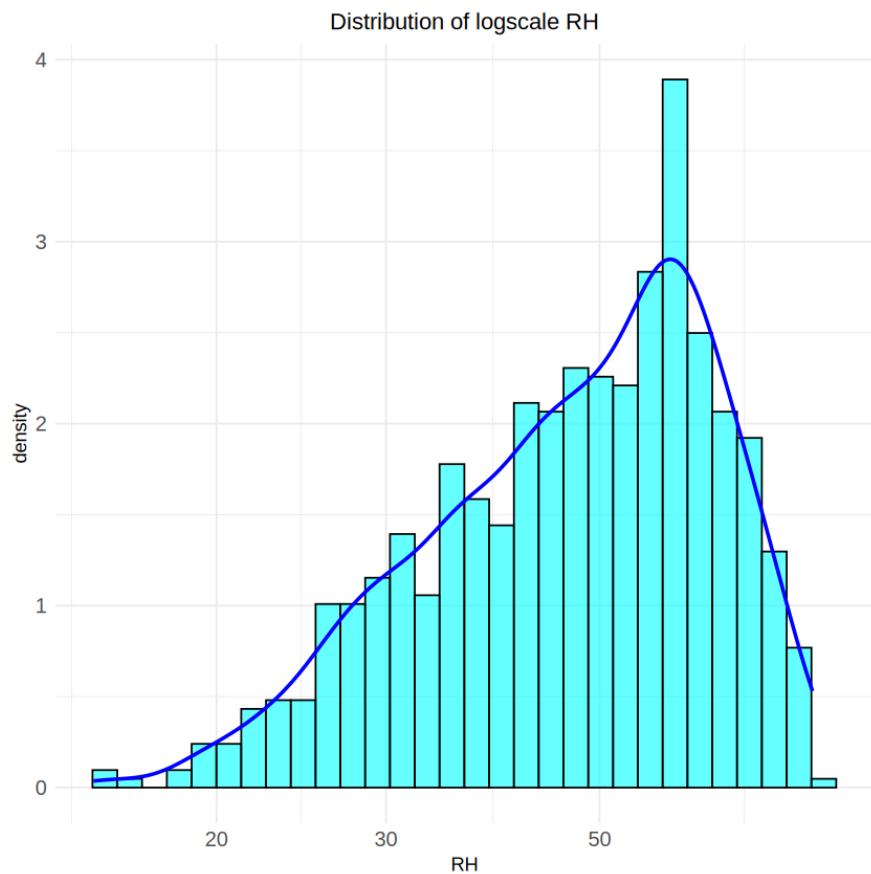


Hình 3.100: Phân phối ban đầu của Sensor response cho Nitrogen oxides.

Nhận xét:

- Nhìn vào histogram trên, ta thấy phân phối của biến này xấp xỉ chuẩn.

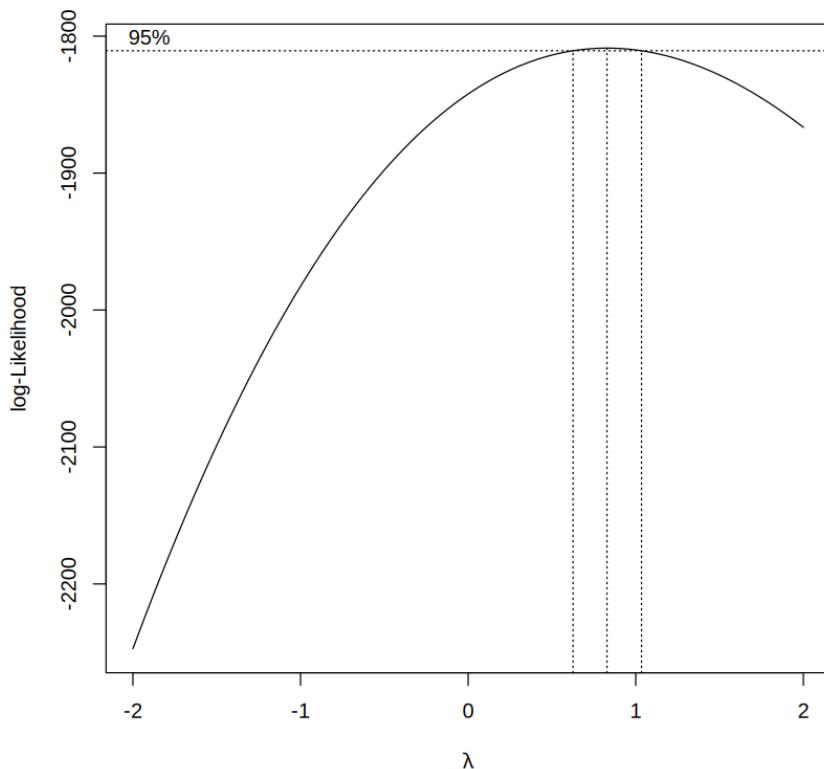
Ta thử sử dụng log-transform nó.



Hình 3.101: Phân phối sau khi log-scale của Sensor response cho Nitrogen oxides.

Nhận xét:

- Sau khi logscale, ta thấy phân phối của biến này có chiều hướng lệch phải (lệch âm)

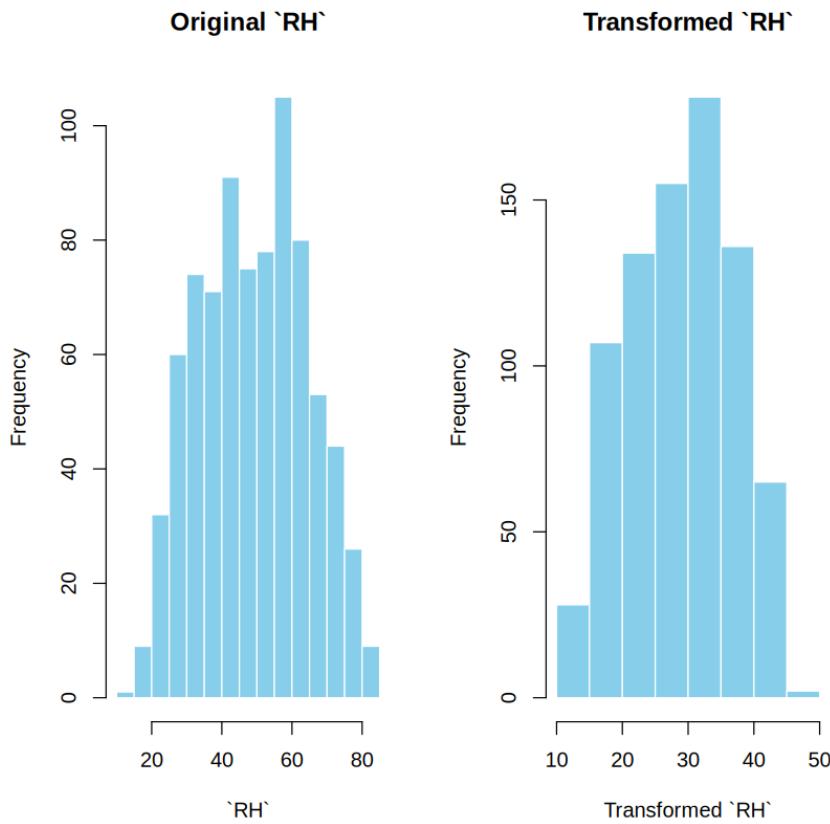


Hình 3.102: Log-likelihood với các giá trị λ của Sensor response cho Nitrogen oxides.

Nhận xét:

- Với mức ý nghĩa 5%, ta tìm được giá trị lambda phù hợp 0.909

Và ta thực hiện biến đổi dữ liệu với giá trị lambda vừa tìm được.

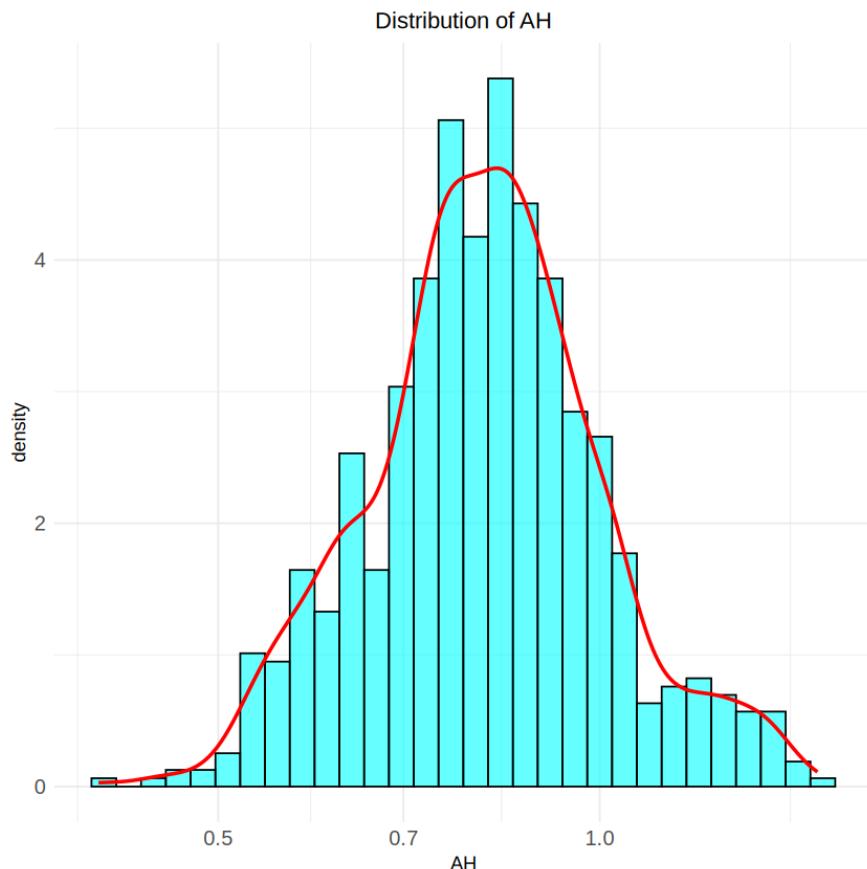


Hình 3.103: Phân phối trước và sau khi biến đổi của Sensor response cho Nitrogen oxides.

Nhận xét:

- Ta có được giá trị lambda tối ưu là 0.909 và sử dụng giá trị này để biến đổi biến RH. Biểu đồ histogram phía bên dưới thể hiện phân phối của biến này trước và sau khi biến đổi. Dễ dàng thấy được, sau khi biến đổi, biến này đã tương đối chuẩn hơn.

AH: Absolute humidity (g/m³)

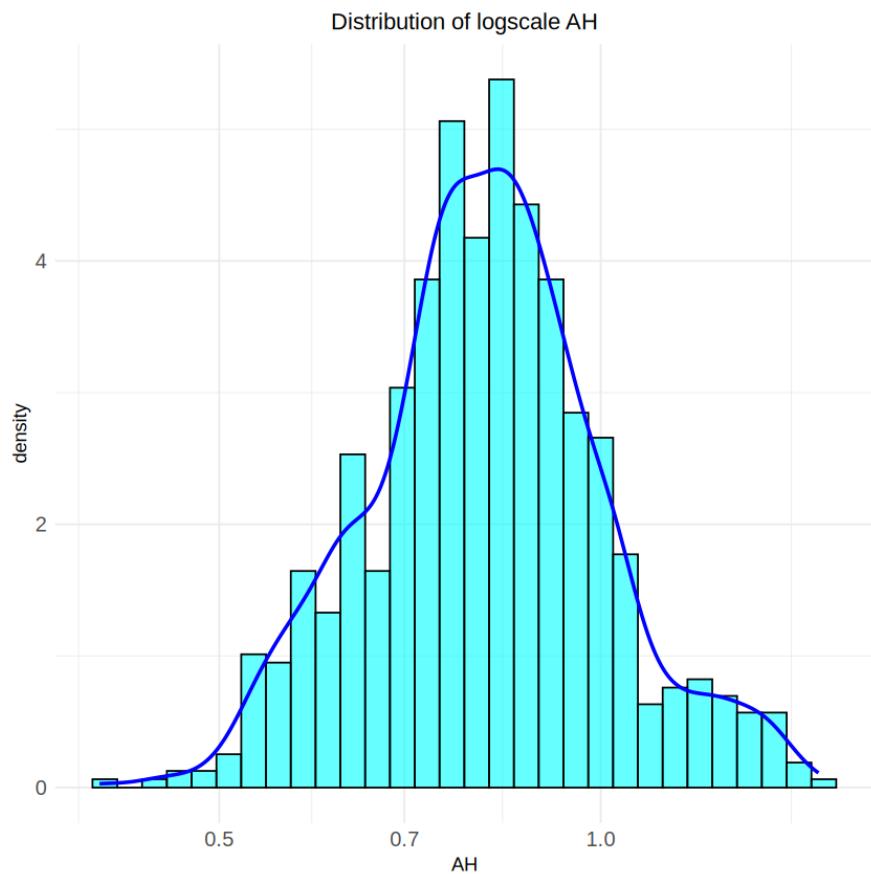


Hình 3.104: Phân phối ban đầu của Sensor response cho Nitrogen oxides.

Nhận xét:

- Nhìn vào histogram trên, ta thấy phân phối của biến bị lệch phải (lệch âm).

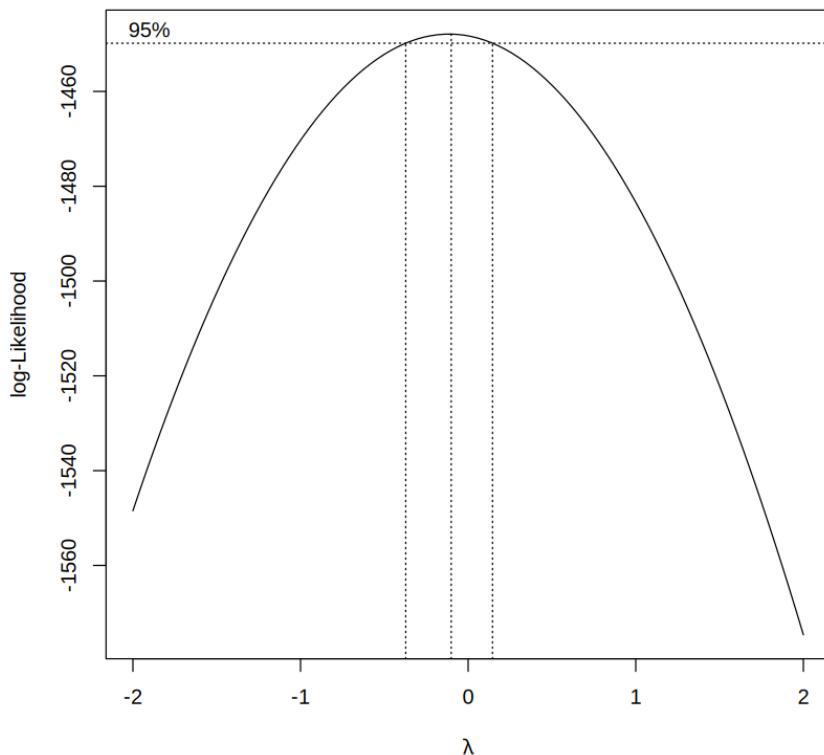
Ta thử sử dụng log-transform nó.



Hình 3.105: Phân phối sau khi log-scale của Sensor response cho Nitrogen oxides.

Nhận xét:

- Sau khi logscale, phân phối của biến này vẫn bị lệch.

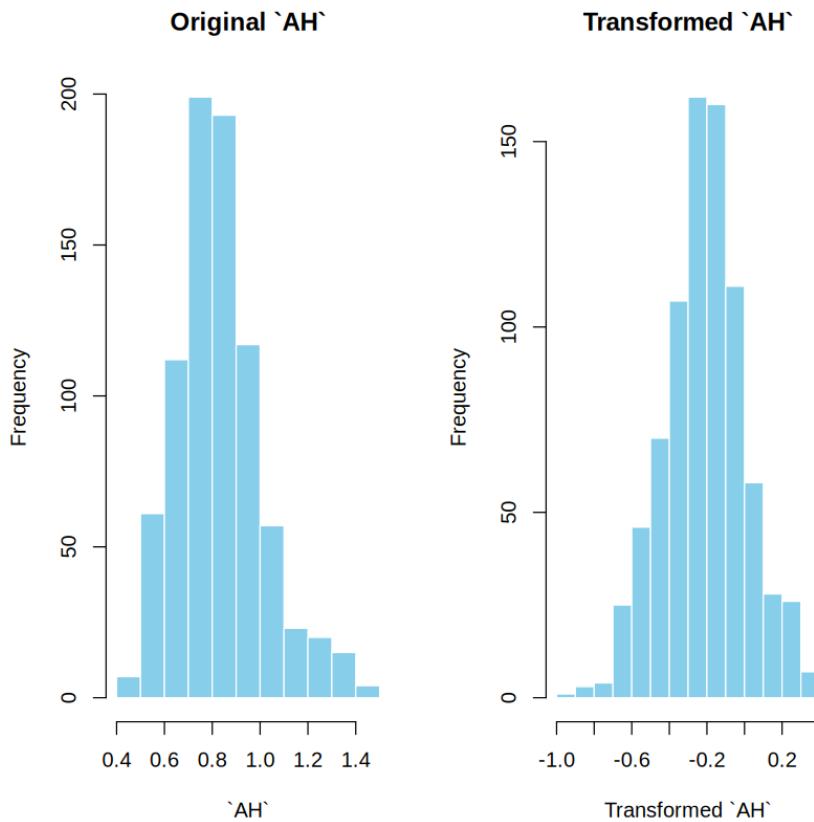


Hình 3.106: Log-likelihood với các giá trị λ của Sensor response cho Nitrogen oxides.

Nhận xét:

- Với mức ý nghĩa 5%, ta tìm được giá trị lambda phù hợp 0.6666

Và ta thực hiện biến đổi dữ liệu với giá trị lambda vừa tìm được.

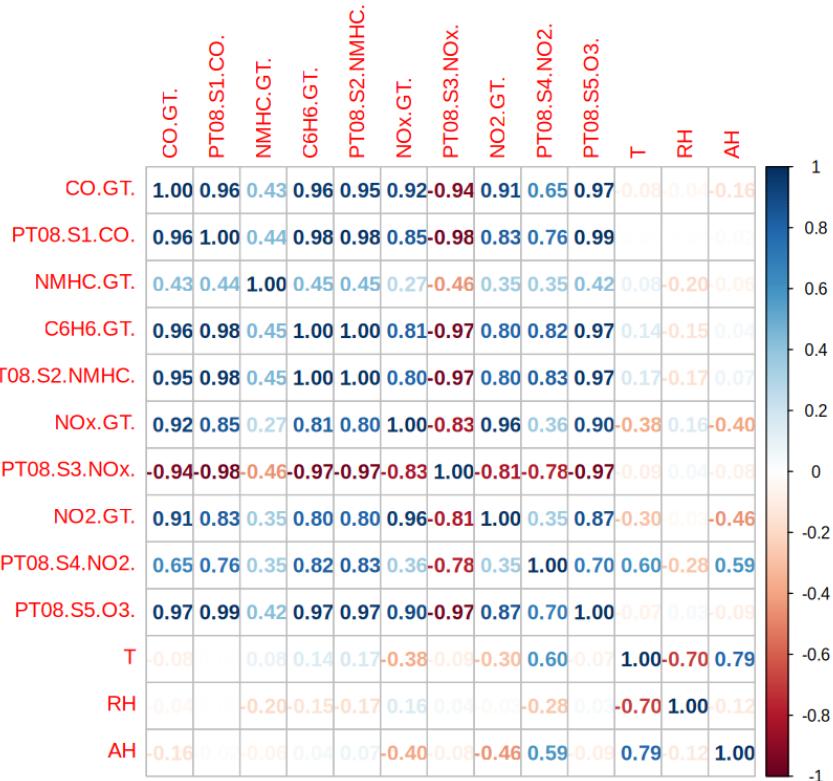


Hình 3.107: Phân phối trước và sau khi biến đổi của Sensor response cho Nitrogen oxides.

Nhận xét:

- Ta có được giá trị lambda tối ưu là 0.6666 và sử dụng giá trị này để biến đổi biến AH. Biểu đồ histogram phía bên dưới thể hiện phân phối của biến này trước và sau khi biến đổi. Dễ dàng thấy được, sau khi biến đổi, biến này đã tương đối chuẩn hơn.
- Ta dễ ý thấy có một số ngoại lệ ở giá trị -0.2.

3.2.6. Phân tích đa biến



Hình 3.108: Ma trận tương quan giữa các biến trong tập dữ liệu Air.

Nhận xét:

- Ta thấy các biến trong tập dữ liệu chất lượng không khí có tương quan mạnh với nhau. Điều này có thể chỉ ra rằng trong tập dữ liệu này có hiện tượng đa cộng tuyến rất cao.

Phân tích cụ thể, ta xem xét các biến có mối tương quan thuận mạnh:

	Var1	Var2	value
1	CO.GT.	PT08.S1.CO.	0.7733943
2	CO.GT.	C6H6.GT.	0.8123923
3	CO.GT.	PT08.S2.NMHC.	0.7955864
4	CO.GT.	NOx.GT.	0.7622972
5	CO.GT.	PT08.S5.O3.	0.7590266
6	PT08.S1.CO.	C6H6.GT.	0.8838209
7	PT08.S1.CO.	PT08.S2.NMHC.	0.8929724
8	PT08.S1.CO.	PT08.S3.NOx.	-0.7719181

10	9	PT08.S1.CO.	PT08.S5.03.	0.8993259
11	10	C6H6.GT.	PT08.S2.NMHC.	0.9819620
12	11	C6H6.GT.	PT08.S3.NOx.	-0.7357112
13	12	C6H6.GT.	PT08.S4.NO2.	0.7657168
14	13	C6H6.GT.	PT08.S5.03.	0.8657266
15	14	PT08.S2.NMHC.	PT08.S3.NOx.	-0.7966873
16	15	PT08.S2.NMHC.	PT08.S4.NO2.	0.7772348
17	16	PT08.S2.NMHC.	PT08.S5.03.	0.8805903
18	17	NOx.GT.	NO2.GT.	0.7631329
19	18	PT08.S3.NOx.	PT08.S5.03.	-0.7965536

Nhận xét:

- Trong dữ liệu, các biến có sự tương quan mạnh với nhau.
- Ta dễ dàng thấy được, đa số cặp tương quan đều có chỉ số cao hơn 0.7
- Điều này giúp chúng ta định hướng được nên sử dụng các mô hình có khả năng xử lý hiện tượng đa cộng tuyến cao trong dữ liệu.

Và ta cũng xem xét các cặp biến có mối tương quan nghịch mạnh

		Var1	Var2	value
2	1	CO.GT.	PT08.S3.NOx.	-0.61387028
3	2	PT08.S1.CO.	PT08.S3.NOx.	-0.77191812
4	3	NMHC.GT.	PT08.S3.NOx.	-0.26198519
5	4	NMHC.GT.	RH	-0.05279411
6	5	C6H6.GT.	PT08.S3.NOx.	-0.73571121
7	6	C6H6.GT.	RH	-0.06164347
8	7	PT08.S2.NMHC.	PT08.S3.NOx.	-0.79668731
9	8	PT08.S2.NMHC.	RH	-0.09035172
10	9	NOx.GT.	PT08.S3.NOx.	-0.56325927
11	10	NOx.GT.	T	-0.23565652
12	11	NOx.GT.	AH	-0.12683093
13	12	PT08.S3.NOx.	NO2.GT.	-0.56953458
14	13	PT08.S3.NOx.	PT08.S4.NO2.	-0.53846012
15	14	PT08.S3.NOx.	PT08.S5.03.	-0.79655364
16	15	PT08.S3.NOx.	T	-0.14513301
17	16	PT08.S3.NOx.	RH	-0.05672984
18	17	PT08.S3.NOx.	AH	-0.23202063
19	18	NO2.GT.	T	-0.16531710

```

20 19      NO2.GT.          RH -0.08064472
21 20      NO2.GT.          AH -0.29119971
22 21 PT08.S4.NO2.          RH -0.03218809
23 22 PT08.S5.03.           T -0.02719336
24 23             T          RH -0.57856879

```

Nhận xét:

- Chúng ta cũng rút ra kết luận tương tự.

Khảo sát ngoại lai

Ta sử dụng IQR để tìm các điểm ngoại lai và cực ngoại lai:

- Tổng số ngoại lai: 0
- Tổng số cực ngoại lai: 0

Chuẩn hóa và phân chia tập dữ liệu

Ta sử dụng box-cox transform và sau đó phân chia tập dữ liệu thành 2 phần: train (70%) và test (30%).

3.2.7. Mô hình hóa bằng hồi quy tuyến tính đa biến

Trước tiên, ta xây dựng mô hình đầy đủ như sau

```

1 full.lm <- lm(`C6H6.GT.` ~ ., data = train)
2 print(summary(full.lm))

```

Kết quả:

```

1 lm(formula = C6H6.GT. ~ ., data = train)
2
3 Residuals:
4   Min     1Q Median     3Q    Max
5 -1.3855 -0.1963 -0.0553  0.1558  3.9380
6
7 Coefficients:
8                   Estimate Std. Error t value Pr(>|t|)
9 (Intercept) -1.423e+01  1.795e-01 -79.253 < 2e-16 ***
10 CO.GT.      4.689e-02  8.772e-03   5.346 9.32e-08 ***
11 PT08.S1.CO. -1.518e-03  4.185e-04  -3.626 0.000290 ***

```

```

12 NMHC.GT.      1.064e-03  2.993e-04   3.556  0.000379 *** 
13 PT08.S2.NMHC. 1.083e-01  5.774e-04  187.655 < 2e-16 *** 
14 NOx.GT.       8.187e-03  2.838e-04  28.845 < 2e-16 *** 
15 PT08.S3.NOx.  1.125e-02  3.328e-04  33.819 < 2e-16 *** 
16 NO2.GT.        -1.848e-02 7.815e-04 -23.649 < 2e-16 *** 
17 PT08.S4.NO2.  5.558e-03  3.845e-04  14.453 < 2e-16 *** 
18 PT08.S5.03.   1.357e-03  2.494e-04  5.443  5.44e-08 *** 
19 T              -8.847e-02 7.468e-03 -11.846 < 2e-16 *** 
20 RH             -4.064e-02 3.545e-03 -11.464 < 2e-16 *** 
21 AH             5.389e-01  4.896e-02  11.008 < 2e-16 *** 
22 --- 
23 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
24 ... 
25 Residual standard error: 0.3119 on 6536 degrees of freedom 
26 Multiple R-squared:  0.9923 ,    Adjusted R-squared:  0.9923 
27 F-statistic: 7.05e+04 on 12 and 6536 DF,  p-value: < 2.2e-16

```

Nhận xét:

- Mô hình có R-squared là 0.9923, tức là 99.23% phuơng sai của biến C6H6.GT. được giải thích bởi các biến độc lập trong mô hình.

Tiếp theo, ta định nghĩa mô hình chặn trên và chặn dưới để lựa chọn mô hình

```

1 # Mô hình chặn dưới 
2 model.lb <- lm(`C6H6.GT.` ~ 1, data = train) 
3 
4 # Mô hình chặn trên 
5 model.up <- full.lm 
6 
7 step(full.lm, scope = list(lower = model.lb, upper = model.up), 
      direction = "both", trace = FALSE)

```

Kết quả:

```

1 lm(formula = C6H6.GT. ~ CO.GT. + PT08.S1.CO. + NMHC.GT. + PT08. 
2 S2.NMHC. + 
3 NOx.GT. + PT08.S3.NOx. + NO2.GT. + PT08.S4.NO2. + PT08.S5. 
4 O3. + 
5 T + RH + AH, data = train)

```

```

5 Coefficients:
6   (Intercept) CO.GT. PT08.S1.CO. NMHC.GT.
7   PT08.S2.NMHC.
8   -14.229708 0.046891 -0.001518 0.001064
9   0.108347
10  NOx.GT. PT08.S3.NOx. NO2.GT. PT08.S4.NO2.
11  PT08.S5.O3.
12  0.008187 0.011255 -0.018483 0.005558
13  0.001357
14  T RH AH
15  -0.088472 -0.040642 0.538945

```

```

1 air_models<- regsubsets(C6H6.GT. ~ CO.GT. + PT08.S1.CO. + NMHC.
2   GT. + PT08.S2.NMHC. +
3   NOx.GT. + PT08.S3.NOx. + NO2.GT. + PT08.S4.NO2. + PT08.S5.
4   O3. +
5   T + RH + AH, data = train)
6 # Lựa chọn mô hình tốt nhất từ reg subsets
7 summary.air$which

```

Ta lựa chọn mô hình tốt nhất dựa trên BIC

```

1 # Tiêu chí chọn mô hình tốt nhất 4: mô hình với BIC nhỏ
2 summary.air$bic
3
4 best_model_index <- which.min(summary.csm$bic)
5 best_model <- summary.csm$which[best_model_index, ]
6 print(best_model)
7 best_vars <- names(best_model[best_model])
8 best_vars <- best_vars[best_vars != "(Intercept)"]
9 print(best_vars)

```

Kết quả:

```

1 (Intercept) CO.GT. PT08.S1.CO. NMHC.GT. PT08.S2.
2   NMHC.
3   TRUE FALSE FALSE FALSE
4   TRUE

```

```

3      NOx.GT.  PT08.S3.NOx.      NO2.GT.  PT08.S4.NO2.  PT08.
4          S5.03.
5      TRUE        TRUE        TRUE        TRUE
6          FALSE
7      T          RH          AH
8      TRUE        TRUE        TRUE
9 [1] "PT08.S2.NMHC." "NOx.GT."      "PT08.S3.NOx." "NO2.GT."
10 [5] "PT08.S4.NO2." "T"           "RH"          "AH"

```

Xây dựng mô hình tốt nhất

```

1 # Xây dựng mô hình tốt nhất
2 formula_str <- paste("C6H6.GT. ~", paste(best_vars, collapse =
3   " + "))
4
5 # Tóm tắt mô hình
6 summary(best_model_air)

```

Kết quả:

```

1 lm(formula = as.formula(formula_str), data = train)
2
3 Residuals:
4   Min     1Q Median     3Q    Max
5 -1.3449 -0.1997 -0.0546  0.1609  3.8753
6
7 Coefficients:
8                 Estimate Std. Error t value Pr(>|t|)
9 (Intercept) -1.434e+01  1.717e-01 -83.53  <2e-16 ***
10 PT08.S2.NMHC. 1.091e-01  5.156e-04 211.67  <2e-16 ***
11 NOx.GT.       8.707e-03  2.648e-04  32.88  <2e-16 ***
12 PT08.S3.NOx.  1.081e-02  3.248e-04  33.27  <2e-16 ***
13 NO2.GT.      -1.743e-02  7.717e-04 -22.59  <2e-16 ***
14 PT08.S4.NO2.  5.917e-03  3.628e-04  16.31  <2e-16 ***
15 T            -9.191e-02  7.381e-03 -12.45  <2e-16 ***
16 RH           -4.096e-02  3.558e-03 -11.51  <2e-16 ***
17 AH           5.359e-01  4.887e-02  10.97  <2e-16 ***
18 ---
19 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

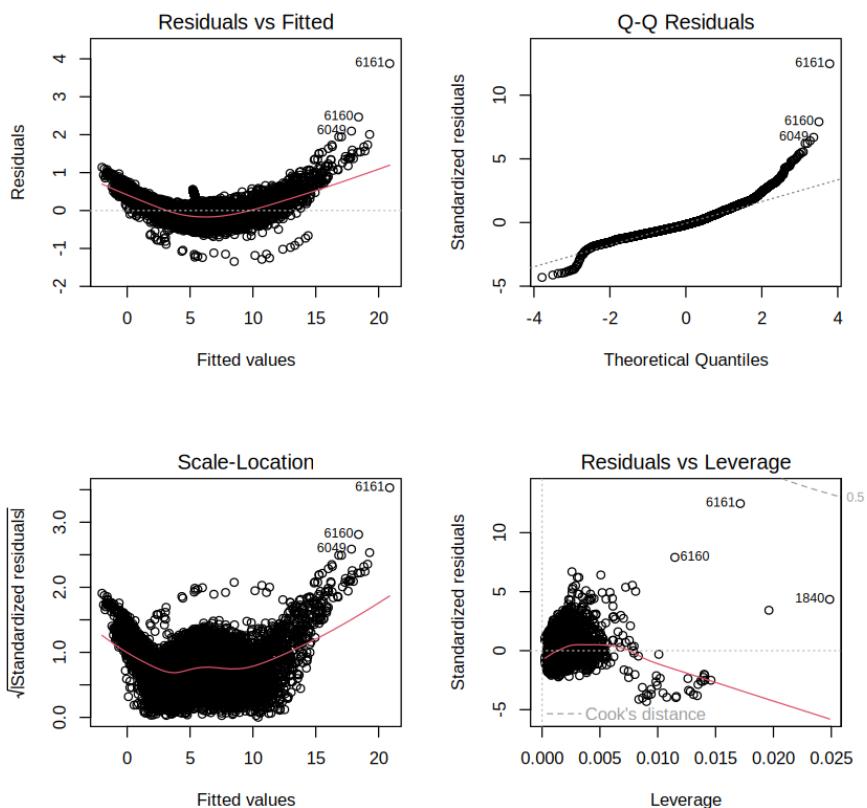
20
21 Residual standard error: 0.3136 on 6540 degrees of freedom
22 Multiple R-squared:  0.9922 ,   Adjusted R-squared:  0.9922
23 F-statistic: 1.046e+05 on 8 and 6540 DF,  p-value: < 2.2e-16

```

Nhận xét:

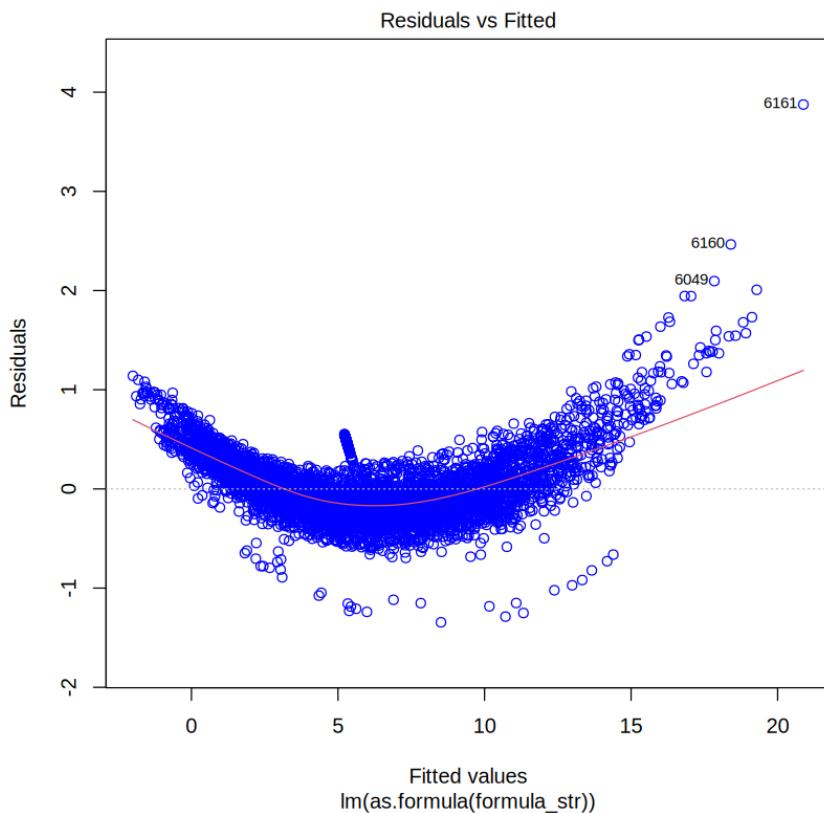
- Mô hình có R-squared là 0.9922, tức là 99.22% phương sai của biến C6H6.GT. được giải thích bởi các biến độc lập trong mô hình.

Bây giờ, ta sẽ đi phân tích mô hình này.



Hình 3.109: .

Phân tích Residuals vs Fitted Plot: Biểu đồ Residuals vs Fitted Plot đưa ra dấu hiệu nếu có các mẫu phi tuyến tính. Để hồi quy tuyến tính chính xác, dữ liệu cần phải tuyến tính nên điều này sẽ kiểm tra xem điều kiện đó có được đáp ứng hay không.

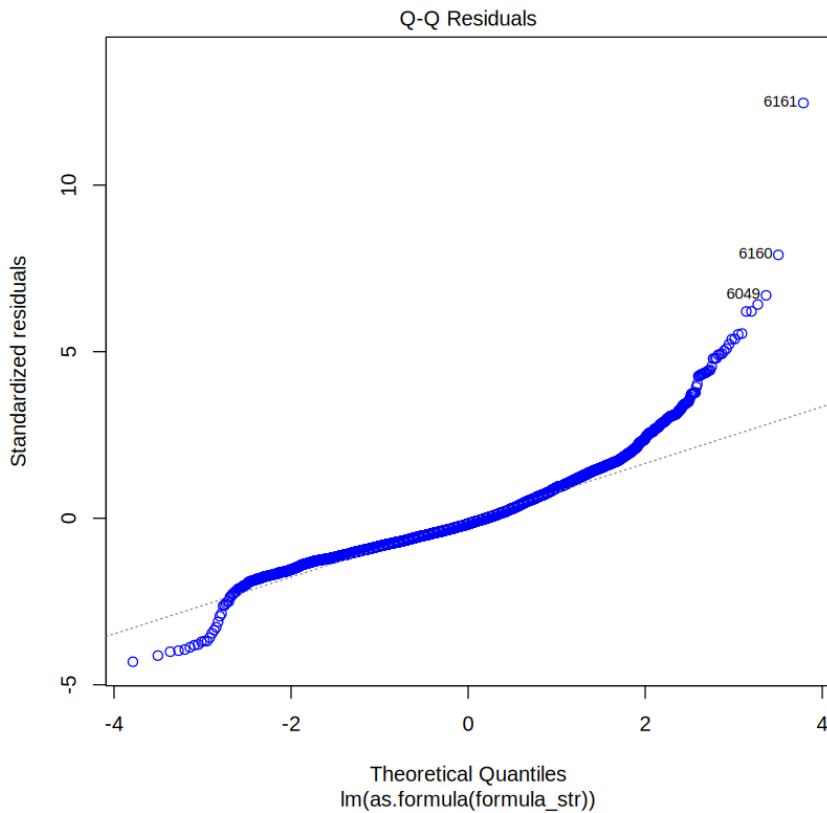


Hình 3.110: Biểu đồ Residuals vs Fitted Plot của mô hình hồi quy Air.

Nhận xét:

- Dựa trên biểu đồ này, ta thấy đường cong màu đỏ có dáng gần như một đường thẳng, và các phần tử trải dọc theo đường cong này một cách tương đồng đều. Điều này chứng tỏ không có quan hệ phi tuyến xuất hiện trong dữ liệu.

Phân tích Normal Q–Q (quantile-quantile) Plot: Các giá trị thặng dư (residual) nên có phân phối chuẩn. Để kiểm tra điều này, chúng ta cần quan sát biểu đồ QQ Residuals plot, nếu các điểm được xếp thành một đường thẳng (hoặc gần như thẳng) thì chứng tỏ các giá trị thặng dư (residual) có phân phối chuẩn.



Hình 3.111: Normal Q–Q (quantile-quantile) Plot cho mô hình hồi quy Air.

Cẩn thận hơn, chúng ta thử dùng Shapiro–Wilk test để kiểm tra có đúng thật là các giá trị thặng dư có phân phối chuẩn hay không?

- H0: Biến thặng dư của mô hình phân phối chuẩn trong một số quần thể.
- H1: Biến thặng dư của mô hình không phân phối chuẩn trong một số quần thể.

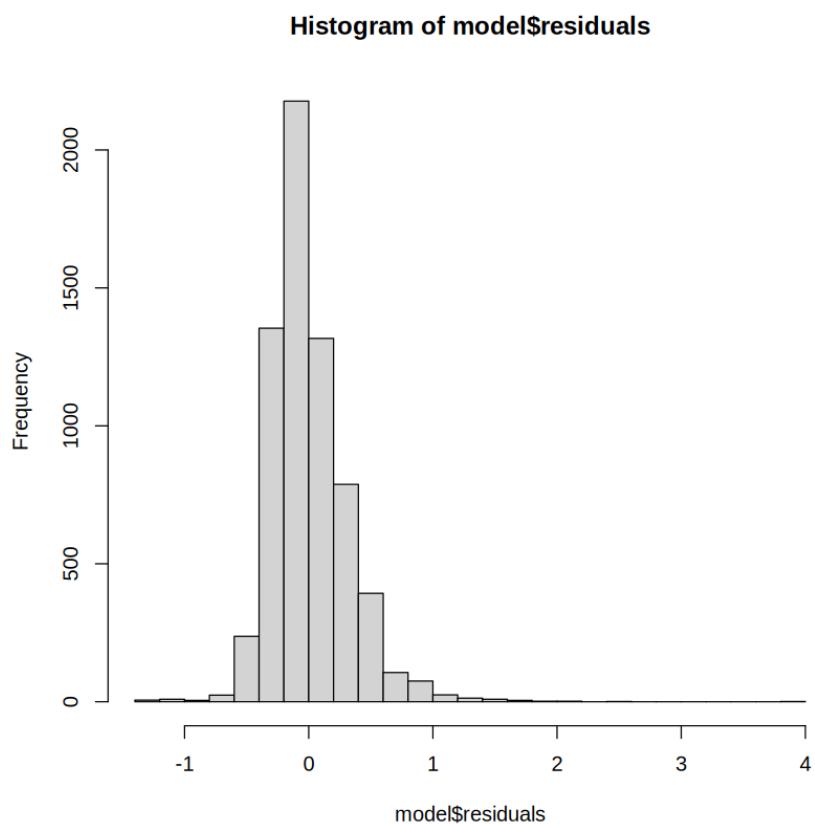
Kết quả:

```

1 Shapiro-Wilk normality test
2
3 data: model$residuals[3:5000]
4 W = 0.96092, p-value < 2.2e-16
5
6 [1] "H0 rejected: the residuals are NOT distributed normally"

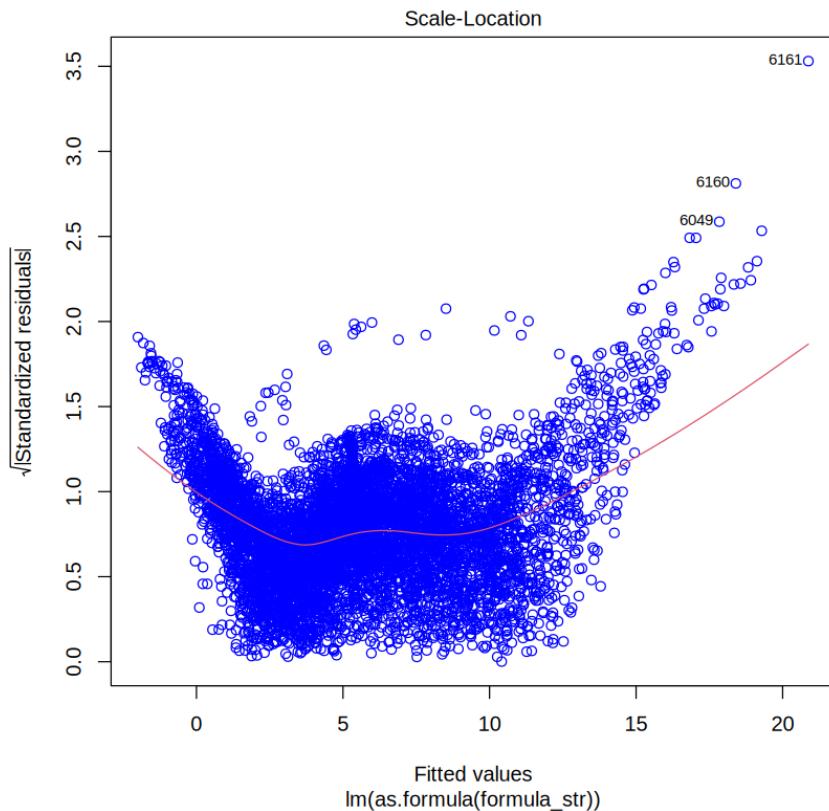
```

Kết quả cho thấy p-value bé hơn mức ý nghĩa alpha 0.05 nên ta có thể bác bỏ giả thuyết H0, biến thặng dư của chúng ta không chuẩn trong một số quần thể. Do biến thặng dư không thỏa mãn được tính chuẩn nên những phân tích về sau có thể không đảm bảo độ tin cậy.



Hình 3.112: Histogram biến thặng dư của mô hình hồi quy Air.

Phân tích Scale-Location: Biểu đồ scale-location kiểm định giả định hồi quy về phương sai bằng nhau (homoscedasticity), tức là giá trị thặng dư có phương sai bằng với đường hồi quy.



Hình 3.113: Scale-Location Plot cho mô hình hồi quy Air.

Nhận xét:

- Đường màu đỏ gần bị lệch về phía dưới của biểu đồ. Nghĩa là, độ phân tán của giá trị thặng dư gần không bằng nhau ở tất cả các giá trị phù hợp.
- Các giá trị thặng dư được phân tán ngẫu nhiên xung quanh đường màu đỏ với độ biến thiên tương đối bằng nhau ở tất cả các giá trị phù hợp.

Cẩn thận hơn, chúng ta sử dụng Breusch-Pagan test để kiểm tra có thật là như vậy không?

- H0: Các giá trị thặng dư là homoscedastic
- H1: Các giá trị thặng dư là heteroscedastic

Kết quả:

```

1 studentized Breusch-Pagan test
2
3 data: model
4 BP = 1157, df = 8, p-value < 2.2e-16

```

```

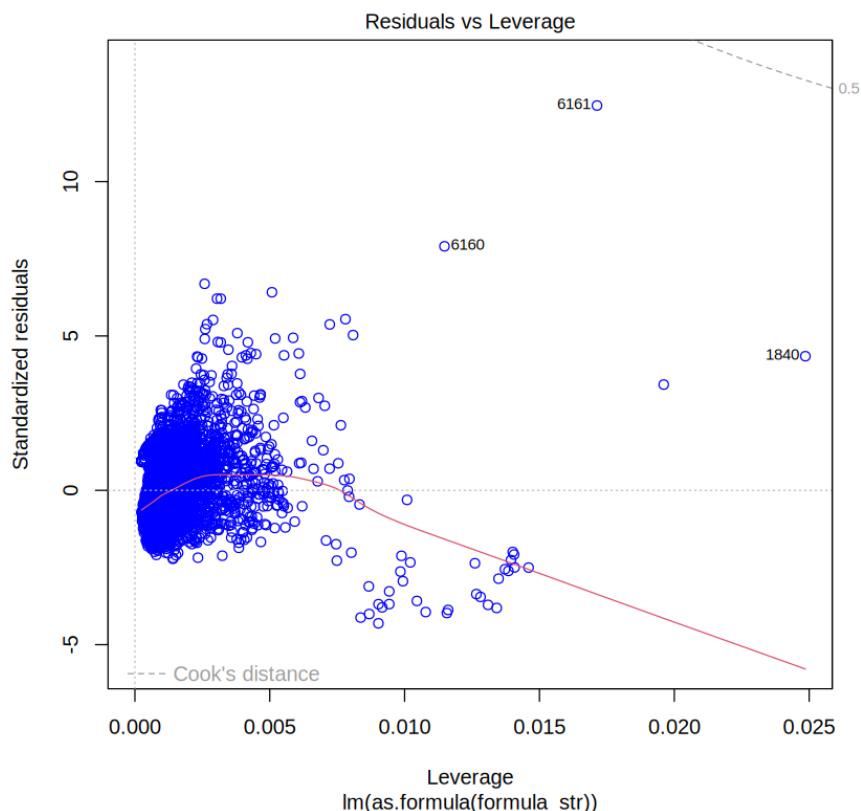
5
6 [1] "H0 rejected: Error variance spreads INCONSTANTLY/
     generating patterns (Heteroscedasticity)"

```

Như vậy, ta thấy p-value lớn hơn mức ý nghĩa 0.05, ta đủ điều kiện bác bỏ H0. Vậy các giá trị thặng dư là heteroscedasticity.

Phân tích Residuals vs Leverage Biểu đồ này có thể được sử dụng để tìm các trường hợp có ảnh hưởng trong tập dữ liệu. Một trường hợp có ảnh hưởng là một trường hợp mà nếu bị loại bỏ sẽ ảnh hưởng đến mô hình nên việc đưa vào hoặc loại trừ nó cần được xem xét.

Một trường hợp có ảnh hưởng có thể là một trường hợp ngoại lệ hoặc không và mục đích của biểu đồ này là xác định các trường hợp có ảnh hưởng lớn đến mô hình. Các ngoại lệ sẽ có xu hướng có giá trị cực cao hoặc cực thấp và do đó ảnh hưởng đến mô hình.

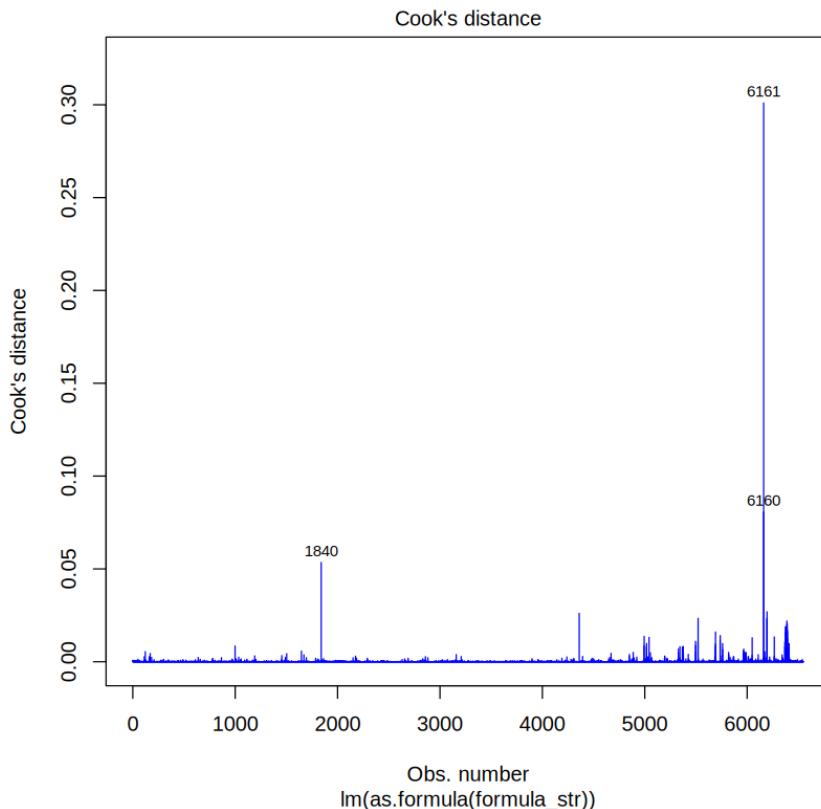


Hình 3.114: Residuals vs Leverage Plot cho mô hình hồi quy Air.

Nhận xét:

- Một số điểm như 1840, 6101, 6100 có thể là các điểm ngoại lai. Ta có thể thử loại bỏ để nâng cao chất lượng mô hình.

Ta cũng có thể nhận thấy có một số giá trị ngoại lai ở cách xa đường thẳng giữa. Ta có thể xem rõ hơn thông qua histogram của Cook's Distance

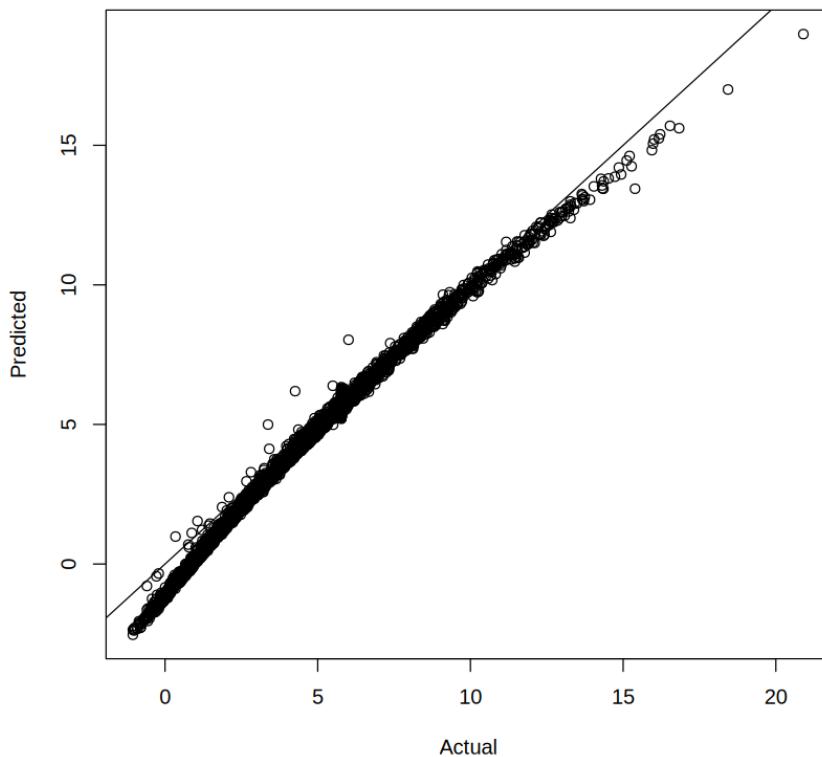


Hình 3.115: Cook Distance Plot của mô hình hồi quy CSM.

Bằng cách loại bỏ các điểm ảnh hưởng dựa trên Cook Distance, ta thử nghiệm việc loại bỏ các giá trị này để xem liệu mô hình có thỏa mãn các điều kiện của mô hình hồi quy tuyến tính hay không. Tuy nhiên, các lần thử nghiệm cho đều cho thấy phân phối của biến thặng dư không chuẩn. Do đó, các phân tích về sau có thể không đảm bảo được tính tin cậy.

Ta trực quan kết quả dự đoán và đánh giá RMSE của mô hình đã chọn.

Predicted vs Actual C6H6(GT) Values (Multiple Regression Model)



Hình 3.116: Trực quan kết quả dự đoán của mô hình Air tốt nhất. RMSE = 0.5201.

Hiệu suất trên các độ đo:

- R-squared: 0.974160108813221
- RMSE: 0.520176239781131

3.2.8. Mô hình hóa bằng PCR

Principal Component Regression (PCR) là một kỹ thuật kết hợp Phân tích thành phần chính (PCA) và hồi quy tuyến tính để giải quyết đa cộng tuyến và giảm chiều trong các tập dữ liệu cao chiều. Các bước chính trong PCR là:

- PCA biến đổi các biến dự báo ban đầu thành một tập hợp các biến mới, không tương quan được gọi là các thành phần chính. Các thành phần này là các tổ hợp tuyến tính của các biến ban đầu và được sắp xếp theo lượng phương sai mà chúng giải thích trong dữ liệu. Mỗi thành phần chính nắm bắt được phương sai tối đa có thể trong khi vẫn trực giao với các thành phần trước đó.
- Một tập hợp con các thành phần chính (giải thích phương sai lớn nhất) được chọn và sử dụng làm các yếu tố dự báo trong mô hình hồi quy tuyến tính để dự báo biến phản hồi. Bằng cách

tập trung vào các thành phần chính nắm bắt được phương sai lớn nhất, PCR hướng đến mục tiêu xây dựng một mô hình hồi quy ổn định và dễ diễn giải hơn.

```
1 # Fitting the PCR model on the training data
2 pcr_model <- pcr(`C6H6.GT.` ~ ., data = train, scale = TRUE,
3   validation = "CV") # Fit PCR model with cross-validation
4 summary(pcr_model)
```

Đối số xác thực = “CV” chỉ định rằng xác thực chéo (cross-validation - CV) nên được sử dụng để xác thực mô hình. Xác thực chéo là một phương pháp mạnh mẽ để đánh giá hiệu suất dự đoán của một mô hình. Nó bao gồm việc phân vùng dữ liệu thành các tập hợp con, huấn luyện mô hình trên một số tập hợp con (bộ huấn luyện - training set) và xác thực nó trên các tập hợp con còn lại (bộ xác thực - validation set). Quá trình này được lặp lại nhiều lần để đảm bảo hiệu suất của mô hình là nhất quán và không phụ thuộc vào phân vùng dữ liệu cụ thể.

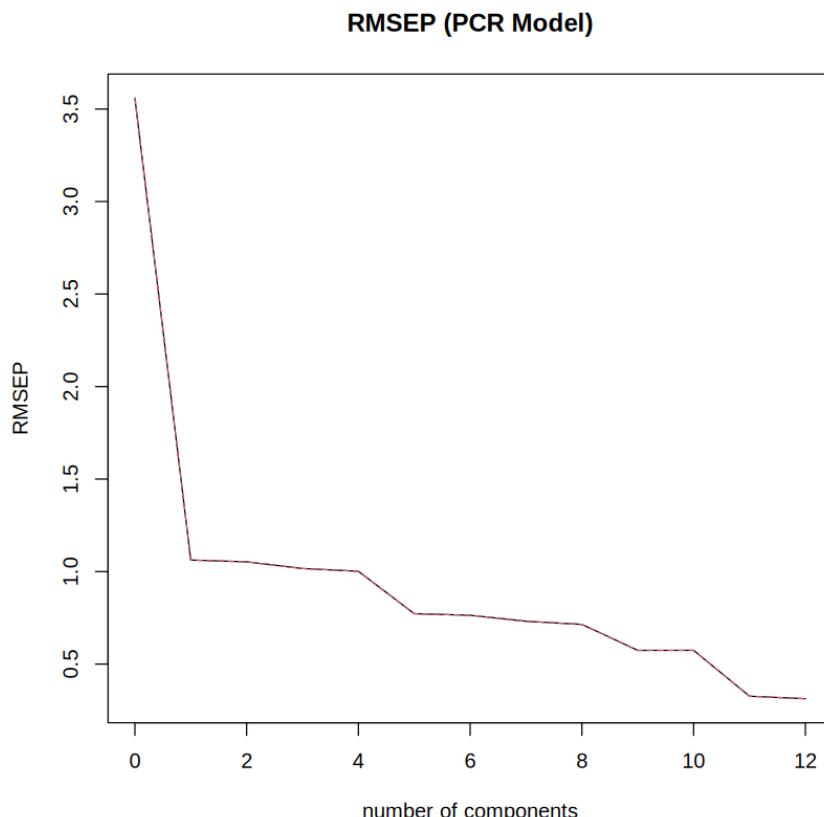
Bằng cách sử dụng xác thực chéo, mô hình ít có khả năng quá khớp với dữ liệu huấn luyện. Quá khớp xảy ra khi mô hình nắm bắt được nhiều và các mẫu cụ thể trong dữ liệu huấn luyện không tổng quát hóa thành dữ liệu mới, chưa từng thấy. Xác thực chéo giúp phát hiện và giảm thiểu tình trạng quá khớp bằng cách kiểm tra mô hình trên các tập hợp con khác nhau của dữ liệu.

Và ta có kết quả như sau:

```
1 Data:    X dimension: 6549 12
2           Y dimension: 6549 1
3 Fit method: svdpc
4 Number of components considered: 12
5
6 VALIDATION: RMSEP
7 Cross-validated using 10 random segments.
8             (Intercept) 1 comps  2 comps  3 comps  4 comps  5 comps
9                 3.559     1.063     1.053     1.016     1.001     0.7721
10                0.7637
11 adjCV          3.559     1.063     1.053     1.016     1.001     0.7721
12                0.7637
13
14                 7 comps  8 comps  9 comps  10 comps 11 comps 12 comps
15 CV            0.7315    0.7142    0.5740    0.5741    0.3266    0.3129
16 adjCV        0.7314    0.7141    0.5738    0.5740    0.3265    0.3128
17
18 TRAINING: % variance explained
```

16		1 comps	2 comps	3 comps	4 comps	5 comps	6 comps
17	X	51.80	68.66	80.52	87.53	92.94	95.34
18		97.08					
C6H6.GT.		91.09	91.26	91.86	92.10	95.30	95.41
		95.79					
19		8 comps	9 comps	10 comps	11 comps	12 comps	
20	X	98.14	98.95	99.65	99.89	100.00	
21	C6H6.GT.	95.99	97.41	97.41	99.17	99.23	

Lựa chọn số lượng thành phần chính: Để quyết định được số thành phần chính tối ưu, chúng ta cần phải trung hòa giữa độ phức tạp của mô hình (tức là số lượng components) và RMSEP (Root Mean Squared Error of Prediction).



Hình 3.117: Giá trị RMSEP với số lượng thành phần chính khác nhau của mô hình Air PCR.

Nhận xét:

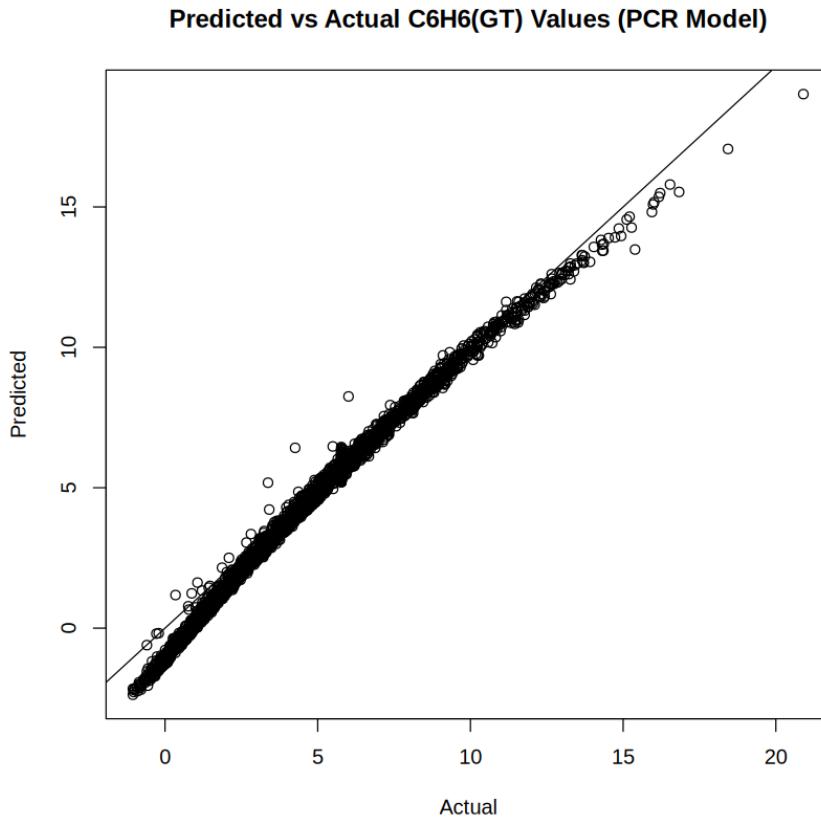
- Dựa trên RMSEP, ta thấy khi sử dụng 12 component mô hình cho kết quả RMSE thấp nhất. Do đó, ta sẽ chọn 12 components

```

1 # Predict using the model and evaluate on the test set with
  optimal number of components
2 optimal_number_of_components <- 12 # Optimal number of
  components based on the RMSEP plot and summary
3 predictions <- predict(pcr_model, ncomp = optimal_number_of_
  components, newdata = test)
4
5 # Compare predictions with actual values
6 plot(test$`C6H6.GT.`, predictions, xlab = "Actual", ylab =
  "Predicted", main = "Predicted vs Actual C6H6(GT) Values (PCR
  Model)") # Plot actual vs predicted values
7 abline(0, 1) # Add a diagonal line for reference

```

Dánh giá kết quả dự đoán



Hình 3.118: Kết quả dự đoán của mô hình Air PCR.

Tính toán RMSE

```

1 # Calculate and print the Root Mean Squared Error (RMSE)
2 rmse <- sqrt(mean((test`C6H6.GT.` - predictions)^2)) # 
    Calculate RMSE between actual and predicted values
3 print(paste("RMSE: ", rmse))

```

Kết quả: RMSE = 0.517225894818149

Tính toán R-squared

```

1 # Calculate the sum of squares of residuals
2 ss_res <- sum((test`C6H6.GT.` - predictions)^2)
3
4 # Calculate the total sum of squares
5 ss_tot <- sum((test`C6H6.GT.` - mean(test`C6H6.GT.))^2)
6
7 # Calculate R-squared
8 r_squared <- 1 - (ss_res / ss_tot)
9
10 # Print R-squared
11 print(paste("R-squared: ", r_squared))

```

Kết quả: R-squared = 0.97445239587654

3.2.9. Mô hình hóa bằng PLS

Partial Least Squares Regression là một kỹ thuật, không giống như PCR, xem xét cả các biến dự báo và biến phản hồi trong quá trình giảm chiều. Các bước chính trong PLS là:

- Latent Variable Extraction: PLS trích xuất một tập hợp các biến tiềm ẩn (thành phần) tối đa hóa hiệp phương sai giữa các biến dự báo và biến phản hồi. Các thành phần này là các tổ hợp tuyến tính của các biến ban đầu, được chọn theo cách mà chúng nắm bắt được càng nhiều thông tin có liên quan càng tốt để dự đoán biến phản hồi. Điều này đảm bảo rằng các thành phần được trích xuất có liên quan trực tiếp đến kết quả quan tâm.
- Regression: Các biến tiềm ẩn sau đó được sử dụng làm biến dự báo trong mô hình hồi quy tuyến tính để dự báo biến phản hồi. Bằng cách kết hợp biến phản hồi vào quy trình trích xuất thành phần, PLS hướng đến mục tiêu cải thiện độ chính xác dự báo của mô hình hồi quy.

```

1 # Fitting the PLS model on the training data
2 pls_model <- plsr(`C6H6.GT.` ~ ., data = train, scale = TRUE,
    validation = "CV")
3

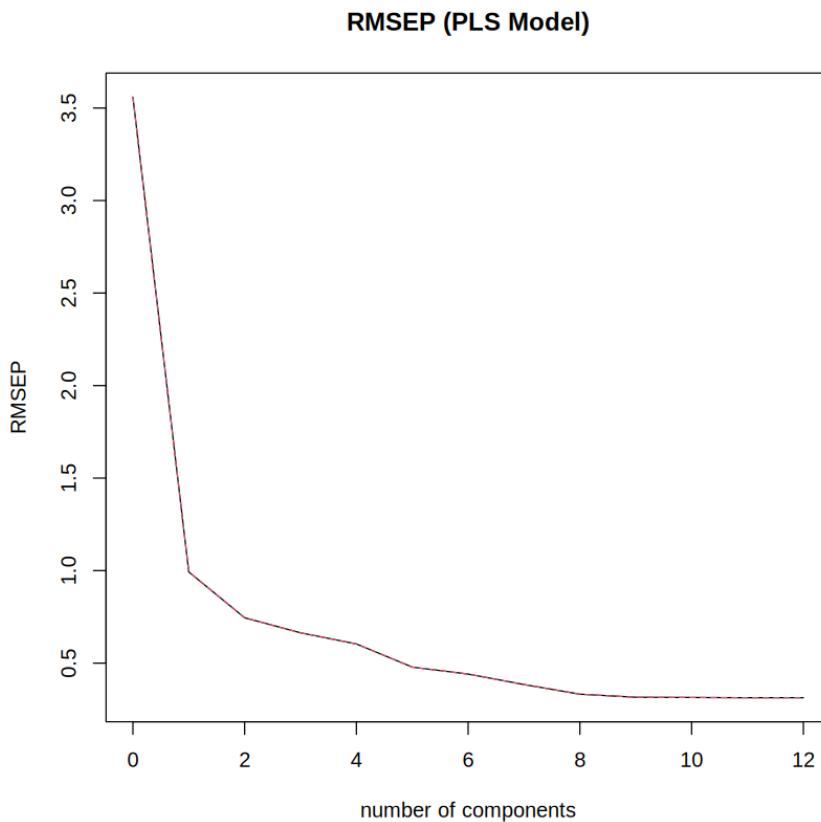
```

```
4 summary(pls_model)
```

Và ta có kết quả:

```
1 Data: X dimension: 6549 12
2 Y dimension: 6549 1
3 Fit method: kernelpls
4 Number of components considered: 12
5
6 VALIDATION: RMSEP
7 Cross-validated using 10 random segments.
8
9          (Intercept) 1 comps 2 comps 3 comps 4 comps 5 comps
10         3.559    0.9928   0.7452   0.6638   0.6035   0.4785
11         0.4408
12 adjCV      3.559    0.9928   0.7450   0.6638   0.6035   0.4783
13         0.4407
14
15 TRAINING: % variance explained
16
17          1 comps 2 comps 3 comps 4 comps 5 comps 6 comps
18          7 comps
19 X        51.76    60.82    72.13    82.96    87.34    92.87
20         95.43
21 C6H6.GT. 92.23    95.63    96.54    97.14    98.21    98.48
22         98.84
23
24          8 comps 9 comps 10 comps 11 comps 12 comps
25 X        96.64    98.12    99.15    99.30    100.00
26 C6H6.GT. 99.14    99.22    99.23    99.23    99.23
```

Lựa chọn số lượng thành phần chính: Để quyết định được số thành phần chính tối ưu, chúng ta cần phải trung hòa giữa độ phức tạp của mô hình (tức là số lượng components) và RMSEP (Root Mean Squared Error of Prediction).



Hình 3.119: Giá trị RMSEP với số lượng thành phần chính khác nhau của mô hình Air PLS.

Nhận xét:

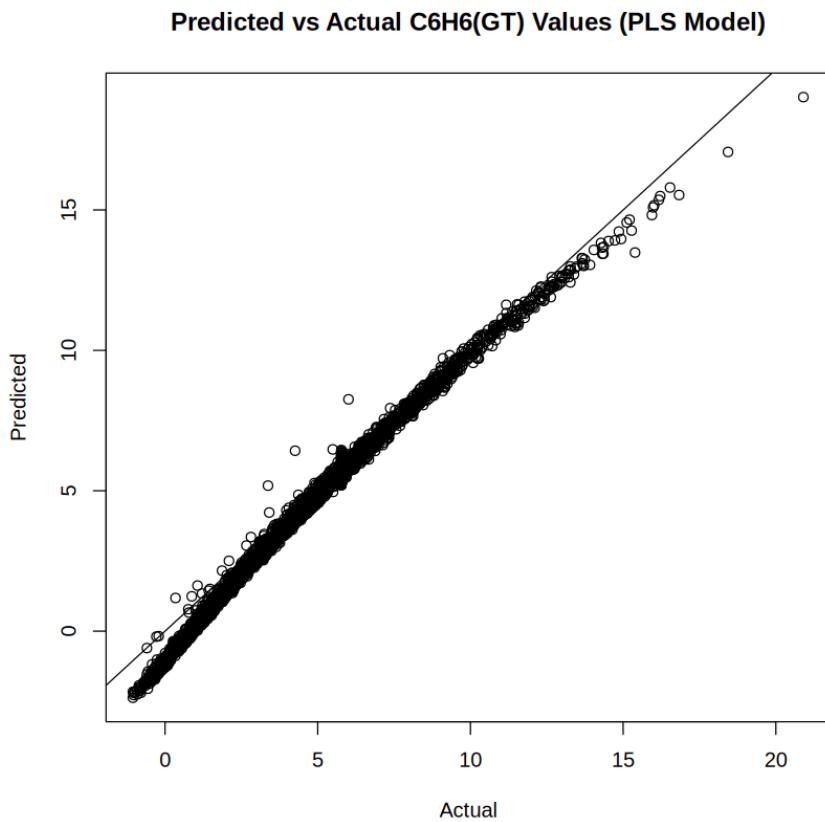
- Dựa trên RMSEP, ta thấy khi sử dụng 11 component mô hình cho kết quả RMSE thấp nhất. Do đó, ta sẽ chọn 11 components

```

1 # Predict using the model and evaluate on the test set with
   optimal number of components
2 optimal_number_of_components <- 11 # Optimal number of
   components based on the RMSEP plot and summary
3 predictions2 <- predict(pls_model, ncomp = optimal_number_of_
   components, newdata = test)
4
5 # Compare predictions with actual values
6 plot(test$`C6H6.GT.`, predictions2, xlab = "Actual", ylab = "
   Predicted", main = "Predicted vs Actual C6H6(GT) Values (PLS
   Model)") # Plot actual vs predicted values
7 abline(0, 1) # Add a diagonal line for reference

```

Đánh giá kết quả dự đoán



Hình 3.120: Kết quả dự đoán của mô hình Air PLS.

Tính toán RMSE

```
1 # Calculate and print the Root Mean Squared Error (RMSE)
2 rmse <- sqrt(mean((test$`C6H6.GT.` - predictions)^2)) # 
   Calculate RMSE between actual and predicted values
3 print(paste("RMSE: ", rmse))
```

Kết quả: RMSE = 0.517942615593671

Tính toán R-squared

```
1 # Calculate the sum of squares of residuals
2 ss_res <- sum((test$`C6H6.GT.` - predictions)^2)
3
4 # Calculate the total sum of squares
5 ss_tot <- sum((test$`C6H6.GT.` - mean(test$`C6H6.GT.`))^2)
6
7 # Calculate R-squared
```

```

8 r_squared <- 1 - (ss_res / ss_tot)
9
10 # Print R-squared
11 print(paste("R-squared: ", r_squared))

```

Kết quả: R-squared = 0.97438154410648

3.2.10. So sánh và đánh giá PCR và PLS

3.2.11. Cải tiến: Random Forest

Xây dựng mô hình Random Forest

```

1 library(randomForest)
2
3 set.seed(123)
4 model_rf <- randomForest(x = train[,-c(4)],
5                             y = train$`C6H6.GT.`,
6                             ntree = 500)
7
8 model_rf

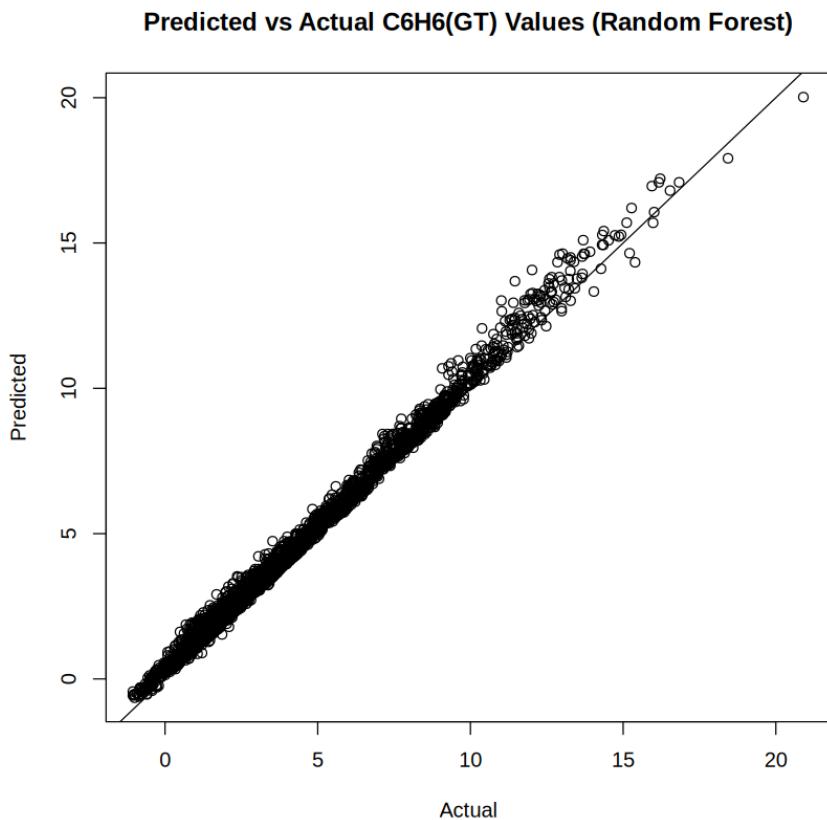
```

Dự đoán và đánh giá kết quả

```

1 library(MLmetrics)
2 library(performance)
3
4 pred_rf_val <- predict(object = model_rf, newdata = test)
5
6 # Compare predictions with actual values
7 plot(test$`C6H6.GT.`, pred_rf_val, xlab = "Actual", ylab = "
     Predicted", main = "Predicted vs Actual C6H6(GT) Values (
     Random Forest)" ) # Plot actual vs predicted values
8 abline(0, 1) # Add a diagonal line for reference

```



Hình 3.121: Kết quả dự đoán của mô hình Air Random Forest.

Đánh giá hiệu suất trên các độ đo

```

1 # Calculate and print the Root Mean Squared Error (RMSE)
2 rmse2 <- sqrt(mean((test$`C6H6.GT.` - pred_rf_val)^2)) #
   Calculate RMSE between actual and predicted values
3 print(paste("RMSE: ", rmse2))

```

Kết quả: RMSE = 0.4141

```

1 # Calculate the sum of squares of residuals
2 ss_res2 <- sum((test$`C6H6.GT.` - pred_rf_val)^2)
3
4 # Calculate the total sum of squares
5 ss_tot2 <- sum((test$`C6H6.GT.` - mean(test$`C6H6.GT.`))^2)
6
7 # Calculate R-squared
8 r_squared2 <- 1 - (ss_res2 / ss_tot2)
9

```

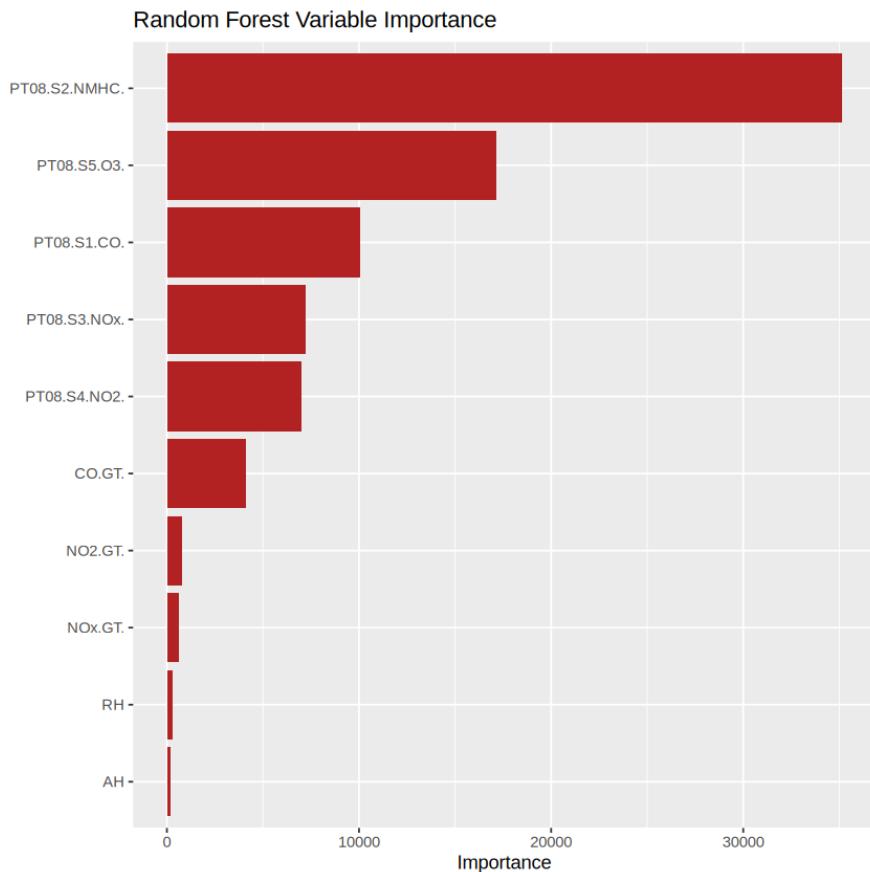
```

10 # Print R-squared
11 print(paste("R-squared: ", r_squared2))

```

Kết quả: R-squared = 0.98362

Mô hình Random Forest có thể giúp ta đánh giá mức độ quan trọng, tức là đóng góp của đặc trưng đối với kết quả của mô hình.



Hình 3.122: Mức độ quan trọng của đặc trưng từ mô hình Air Random Forest.

Nhận xét:

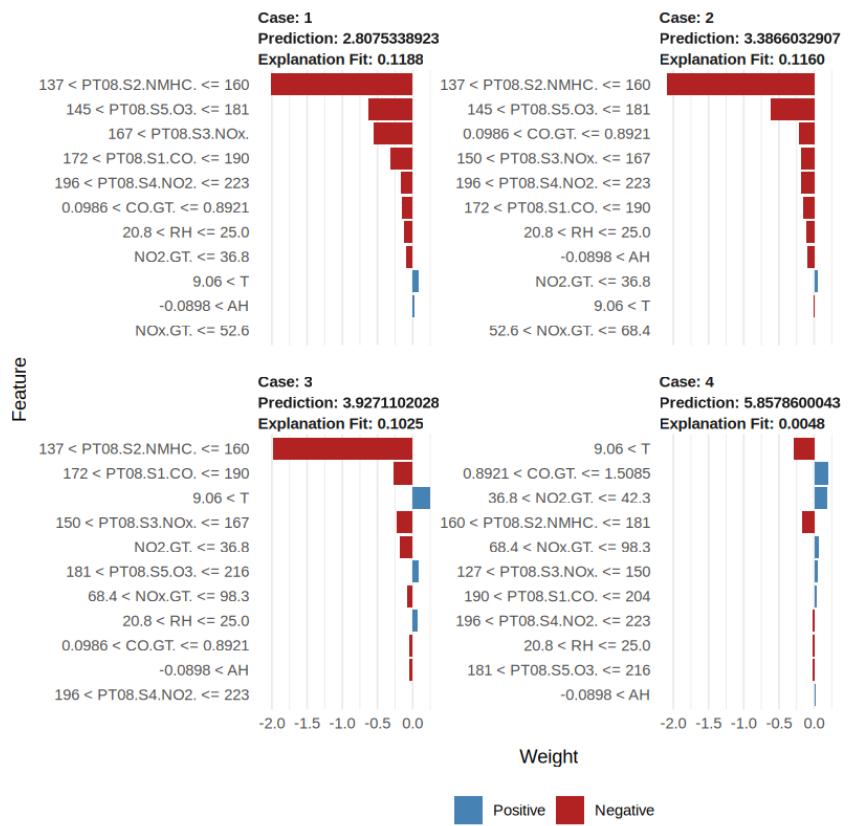
- Ta thấy đặc trưng PT08.S2.NMHC, tức là (Trung bình theo giờ) Phản hồi cảm biến (PPM) của cảm biến hồng ngoại không phân tán (NDIR) đối với Non Metanic HydroCarbons có ảnh hưởng lớn đến kết quả dự đoán C6H6(GT) trong không khí.
- Bên cạnh đó, các đặc tính như CO.GT, PT08.S2.NO2, PT08.S2.NOx, PT08.S2.CO, PT08.S2.O3 cũng có ảnh hưởng đến kết quả dự đoán. Khi so sánh với mô hình hồi quy đa biến, ta cũng nhận thấy các biến này có sự tham gia trong việc giải thích biến C6H6(GT).

Sử dụng phương pháp LIME để giải thích kết quả

```

1 library(lime)
2
3 set.seed(123)
4 explainer <- lime(x = test[,-c(4)],
5                     model = model_rf)
6
7 model_type.randomForest <- function(x){
8   return("regression") # for regression problem
9 }
10
11 predict_model.randomForest <- function(x, newdata, type =
12   "response") {
13
14   # return prediction value
15   predict(x, newdata) %>% as.data.frame()
16
17 }
18
19 # Select only the first 4 observations
20 selected_data <- test[,-c(4)] %>%
21 slice(1:4)
22
23 # Explain the model
24 set.seed(123)
25 explanation <- explain(x = selected_data,
26                         explainer = explainer,
27                         n_features = 27 # Number of features to
28                           explain the model
29 )

```



Hình 3.123: Giải thích kết quả từ mô hình Air Random Forest.

3.2.12. Cải tiến: Support Vector Machine

Xây dựng mô hình SVM

```

1 library(e1071)
2 model_svm <- svm(`C6H6.GT.` ~ ., data = train)
3 pred_svm_val <- predict(object = model_svm, newdata = test)

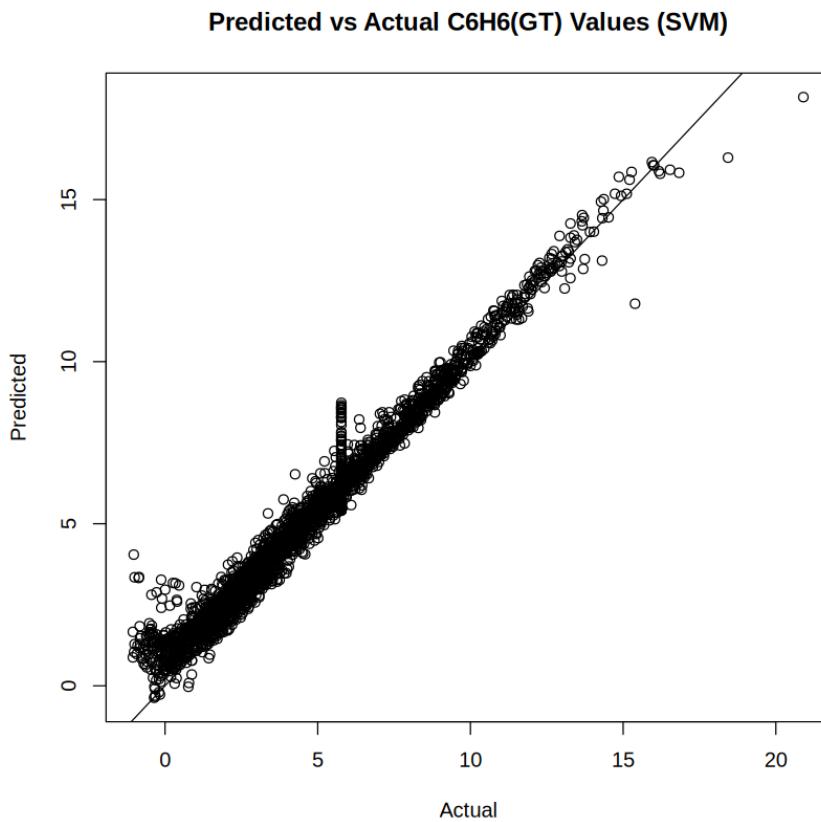
```

Kết quả dự đoán

```

1 # Compare predictions with actual values
2 plot(test$`C6H6.GT.` , pred_svm_val, xlab = "Actual", ylab = "
Predicted", main = "Predicted vs Actual C6H6(GT) Values (SVM
)") # Plot actual vs predicted values
3 abline(0, 1) # Add a diagonal line for reference

```



Hình 3.124: Kết quả dự đoán của mô hình Air SVM.

Đánh giá hiệu suất trên các độ đo

```

1 # Calculate and print the Root Mean Squared Error (RMSE)
2 rmse2 <- sqrt(mean((test$`C6H6.GT.` - pred_svm_val)^2)) #
   Calculate RMSE between actual and predicted values
3 print(paste("RMSE: ", rmse2))

```

Kết quả: RMSE = 0.7319

```

1 # Calculate the sum of squares of residuals
2 ss_res2 <- sum((test$`C6H6.GT.` - pred_svm_val)^2)
3
4 # Calculate the total sum of squares
5 ss_tot2 <- sum((test$`C6H6.GT.` - mean(test$`C6H6.GT.`))^2)
6
7 # Calculate R-squared
8 r_squared2 <- 1 - (ss_res2 / ss_tot2)
9

```

```

10 # Print R-squared
11 print(paste("R-squared: ", r_squared2))

```

Kết quả: R-squared = 0.9488

Sử dụng phương pháp LIME để giải thích kết quả

```

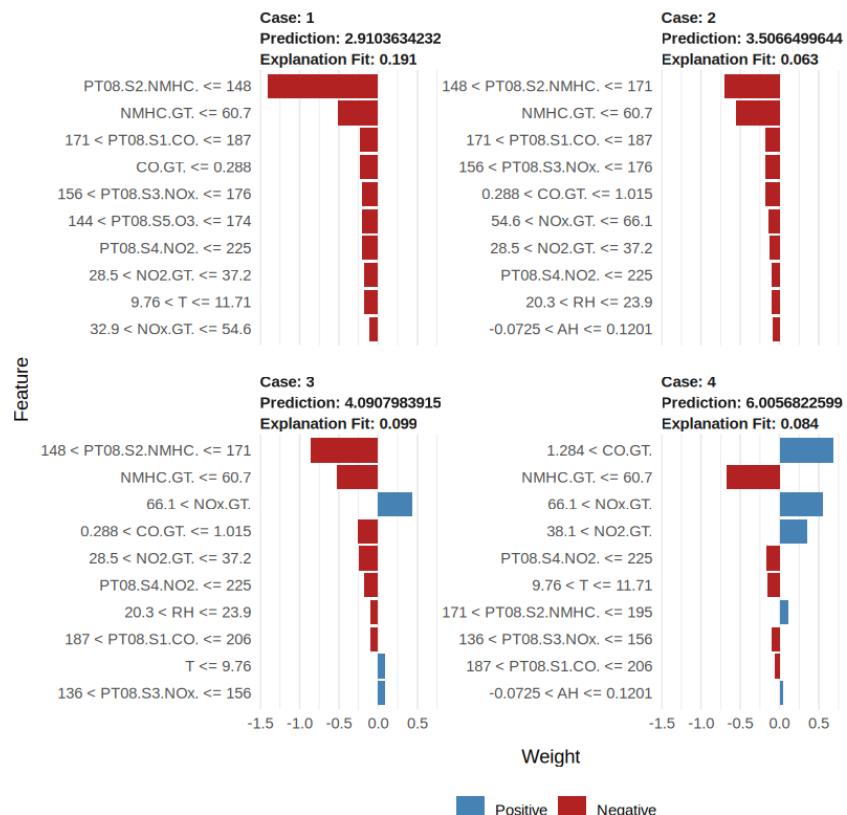
1 library(lime)
2
3 # create the explanation for the SVR model.
4 set.seed(123)
5 explainer_svm <- lime(x = train[,-c(4)],
6                         model = model_svm)
7
8 # Create SVR model specification for lime.
9 model_type.svm <- function(x){
10   return("regression") # for regression problem
11 }
12
13 predict_model.svm <- function(x, newdata, type = "response") {
14
15   # return prediction value
16   predict(x, newdata) %>% as.data.frame()
17
18 }
19
20 set.seed(123)
21 explanation_svm <- explain(x = selected_data,
22                             explainer = explainer_svm,
23                             kernel_width = 1,
24                             feature_select = "auto", # Method of
25                               feature selection for lime
26                             n_features = 10 # Number of features to
27                               explain the model
28 )
29
30 set.seed(123)
31 explanation_svm <- explain(x = selected_data,
32                             explainer = explainer_svm,
33

```

```

32         kernel_width = 1,
33
34         feature_select = "auto", # Method of
35             feature selection for lime
36
37         n_features = 10 # Number of features to
38             explain the model
39
40     )
41
42
43 plot_features(explanation_svm)

```



Hình 3.125: Giải thích kết quả từ mô hình Air SVM.

3.2.13. Kết luận

So sánh hiệu suất của các mô hình dựa trên RMSE

- Mô hình hồi quy đa biến: 0.5201
- Mô hình PCR: 0.5172
- Mô hình PLS: 0.5179

- Mô hình Random Forest: 0.4141
- Mô hình SVM: 0.7319

Và dựa trên R-squared

- Mô hình hồi quy đa biến: 0.9741
- Mô hình PCR: 0.9744
- Mô hình PLS: 0.9743
- Mô hình Random Forest: 0.9836
- Mô hình SVM: 0.9488

Dựa trên kết quả, ta thấy mô hình Random Forest có RMSE thấp nhất. Tuy nhiên, khi sử dụng mô hình này tương đối nặng. Mô hình PCR và PLS cho kết quả cải thiện hơn so với mô hình hồi quy đa biến ban đầu và cũng có tốc độ thực thi phù hợp. Ta xem xét sử dụng mô hình này trong việc dự đoán kết quả.

3.3. Phân tích hiệu quả tương tác của bài đăng trên mạng xã hội facebook

3.3.1. Giới thiệu chung

Trong thời đại kỹ thuật số, mạng xã hội đã trở thành một phần không thể thiếu của cuộc sống hàng ngày. Các nền tảng như Facebook, Twitter, Instagram, LinkedIn, và TikTok không chỉ là nơi để kết nối bạn bè mà còn là môi trường quan trọng để doanh nghiệp tương tác với khách hàng, chính trị gia tiếp cận cử tri, và các nhà hoạt động xã hội lan tỏa thông điệp của họ.

3.3.2. Phát biểu bài toán

Tại sao phân tích tương tác trên mạng xã hội lại quan trọng?

Phân tích tương tác trên mạng xã hội là quá trình thu thập, phân tích và diễn giải các dữ liệu về hành vi của người dùng, bao gồm các lượt thích, chia sẻ, bình luận, và các loại tương tác khác. Bài toán đặt ra là cần hiểu rõ tầm quan trọng của việc phân tích này để từ đó tối ưu hóa các chiến lược truyền thông, tiếp thị, và xây dựng thương hiệu.

Các khía cạnh của bài toán

- **Đo lường hiệu quả chiến lược tiếp thị:** Phân tích tương tác giúp xác định những chiến dịch tiếp thị nào đang hoạt động hiệu quả và đâu là những điểm cần cải thiện. Bằng cách theo dõi các chỉ số như lượt thích, chia sẻ, và bình luận, các nhà tiếp thị có thể điều chỉnh chiến lược để tăng tương tác và thu hút khách hàng.

- **Xây dựng thương hiệu và tăng cường sự hiện diện:** Tương tác cao thường liên quan đến nhận diện thương hiệu mạnh và sự trung thành của khách hàng. Phân tích tương tác giúp xác định những nội dung hoặc chủ đề nào có sức hút đối với người theo dõi, từ đó thúc đẩy việc tạo ra nội dung phù hợp hơn.
- **Hiểu rõ hành vi người dùng:** Phân tích tương tác cung cấp cái nhìn sâu sắc về sở thích và hành vi của người dùng. Hiểu được những gì người dùng thích hoặc không thích có thể giúp doanh nghiệp tạo ra sản phẩm, dịch vụ hoặc nội dung phù hợp hơn với nhu cầu của họ.
- **Phát hiện xu hướng và cơ hội:** Phân tích tương tác cho phép doanh nghiệp và tổ chức phát hiện sớm các xu hướng mới nổi, từ đó tận dụng cơ hội để dẫn đầu thị trường hoặc cộng đồng của mình.
- **Quản lý khủng hoảng:** Khi có khủng hoảng hoặc phản hồi tiêu cực trên mạng xã hội, phân tích tương tác cho phép các tổ chức phát hiện sớm và phản ứng nhanh chóng để giảm thiểu thiệt hại và bảo vệ danh tiếng.

3.3.3. Giới thiệu về tập dữ liệu

Dữ liệu được cho trong tập tin “dataset_Facebook.csv” lấy từ <http://dx.doi.org/10.1016/j.jbusres.2016.02.010>
Thông tin về các biến của bộ dữ liệu như sau:

Variable Name	Role	Type	Description
Page total likes	Feature	Integer	Tổng số lượt thích của bài viết trên trang.
Type	Feature	Categorical	Loại bài viết hoặc quảng cáo (ví dụ: hình ảnh, video, bài viết).
Category	Feature	Integer	Danh mục của bài viết hoặc quảng cáo (có thể là mã số của các loại danh mục khác nhau).
Post Month	Feature	Integer	Tháng mà bài viết được đăng.
Post Weekday	Feature	Integer	Ngày trong tuần mà bài viết được đăng (0 = Chủ Nhật, 1 = Thứ Hai, v.v.).
Post Hour	Feature	Integer	Giờ trong ngày mà bài viết được đăng (0-23).
Paid	Feature	Continuous	Chi phí quảng cáo cho bài viết hoặc quảng cáo (có thể là chi phí phải trả hoặc số tiền đã chi).
Lifetime Post Total Reach	Feature	Integer	Tổng số người tiếp cận bài viết trong suốt thời gian bài viết được đăng.
Lifetime Post Total Impressions	Feature	Integer	Tổng số lần bài viết được nhìn thấy trong suốt thời gian bài viết được đăng.
Lifetime Engaged Users	Feature	Integer	Tổng số người đã tương tác với bài viết (như lượt thích, bình luận, chia sẻ) trong suốt thời gian.

Bảng 3.2: Danh sách các biến và thông tin tương ứng

Trong đó, các biến Post, Lifetime sẽ được chia thành các phần nhỏ hơn (có giải thích chi tiết ở trên). Có tất cả 500 samples.

3.3.4. Đọc và phân tích dữ liệu

Ở bước này, chúng ta sẽ thực hiện một số công việc chính như sau:

- 1 . Đọc dữ liệu và nhận xét tổng quan
- 2 . Thực hiện kiểm tra về bộ dữ liệu bao gồm: Kiểm tra tính độc lập, Kiểm tra dữ liệu khuyết, và kiểm tra outliers của bộ dữ liệu.
- 4 . Trực quan hóa dữ liệu và rút ra nhận xét.

Ngôn ngữ được sử dụng xuyên suốt trong toàn bộ bài báo cáo là R.

Bước 1 : Đọc dữ liệu và nhận xét tổng quan

```

1 data_path = "dataset_Facebook.csv"
2 fb_raw = read.csv(data_path, header = TRUE, sep = ";",
3                   stringsAsFactors = FALSE)
4
5 # Kiểm tra thông tin về tên biến và kiểu dữ liệu
6 str(fb_raw)
7
8 dim(fb_raw)

```

Kết quả trả về như sau:

```

'data.frame': 500 obs. of 19 variables:
 $ Page.total.likes : int 139441 139441 139441 139441 139441 139441 139441 ...
 $ Type : chr "Photo" "Status" "Photo" "Photo" ...
 $ Category : int 2 2 2 2 3 3 2 3 ...
 $ Post.Month : int 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 ...
 $ Post.Weekday : int 3 3 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 ...
 $ Post.Hour : int 3 10 3 10 3 9 3 9 3 10 ...
 $ Paid : int 0 0 0 1 0 0 1 1 0 0 ...
 $ Lifetime.Post.Total.Reach : int 2782 10460 2413 50128 7244 10472 11692 13720 11844 4694 ...
 $ Lifetime.Post.Total.Impressions : int 5091 19057 4973 87991 13594 20849 19479 24137 22538 8668 ...
 $ Lifetime.Engaged.Users : int 178 1457 177 2211 671 1191 481 537 1580 280 ...
 $ Lifetime.Post.Consumers : int 109 1361 119 790 410 1073 265 232 1407 181 ...
 $ Lifetime.Post.Complaints : int 108 117 117 100 100 104 104 104 104 100 ...
 $ Lifetime.Post.Impressions.By.people.who.have.Liked.your.Page : int 3078 31710 2812 61027 6228 16041 15452 19758 15220 4309 ...
 $ Lifetime.Post.reach.By.people.who.like.your.Page : int 640 4012 1503 32048 3200 7852 9328 11056 7912 2329 ...
 $ Lifetime.People.who.have.Liked.your.Page.and.engaged.with.your.post: int 119 1108 132 1386 396 1016 379 422 1250 199 ...
 $ comment : int 4 5 0 58 19 1 3 0 0 3 ...
 $ like : int 79 130 66 1572 325 152 249 325 161 113 ...
 $ share : int 17 19 14 147 49 33 27 14 31 26 ...
 $ Total.Interactions : int 100 164 80 1777 393 186 279 339 192 141 ...
500 19

```

Hình 3.126: Tóm tắt dữ liệu Facebook

Bước 2 : Thực hiện kiểm tra về bộ dữ liệu bao gồm: Kiểm tra tính độc lập, Kiểm tra dữ liệu khuyết, và kiểm tra outliers của bộ dữ liệu.

Hiện tại, bộ dữ liệu chứa rất nhiều biến thông tin, trong phạm vi của đề tài này, chúng tôi sẽ không tiến hành khảo sát tất cả các biến mà chỉ chọn ra 2 biến độc lập để phân tích ANOVA đó là ‘Category’ và ‘Paid’ có ảnh hưởng đến kết quả số lượt thích của bài viết ‘like’ như thế nào.

- Kiểm tra tính độc lập của dữ liệu

```

1 # Kiểm tra tính độc lập của dữ liệu
2 processed_data = fb_raw[c("Category", "Paid", "like")]
3 duplicates = fb_raw[duplicated(processed_data), ]
4 duplicate_counts = table(processed_data[duplicated(
5   processed_data), ])

```

Thực thi đoạn mã trên, ta thấy rằng đối với bộ dữ liệu này, có sự trùng lặp giữa các quan trắc, vậy chúng độc lập với nhau. Vì vậy chúng ta sẽ loại bỏ các điểm này

```
1 rm_duplicated_data = processed_data[!duplicated(  
    processed_data),]  
2 processed_data = rm_duplicated_data  
3 dim(processed_data)
```

Sau bước này, ta thu được 407 samples.

- Kiểm tra dữ liệu khuyết

```
1 missing_ratio = function(s) {  
2   round(mean(is.na(s)) * 100, 1)  
3 }  
4 sapply(islander_raw, missing_ratio)
```

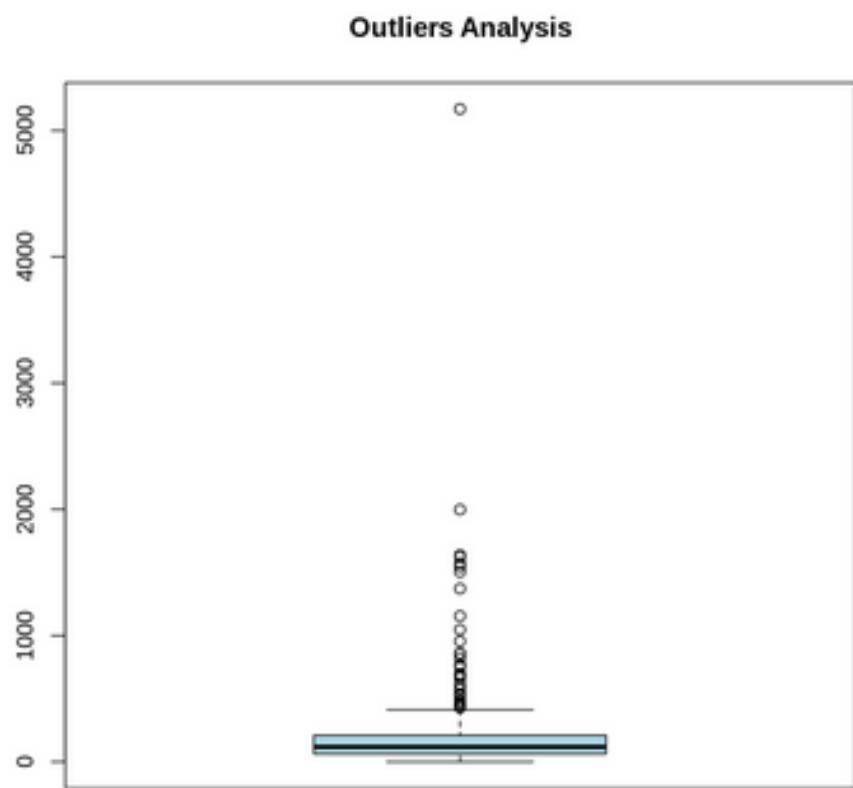
Thực thi đoạn mã trên, ta thấy rằng đối với bộ dữ liệu này, các quan trắc có khuyết đặc trưng, ta tiến hành loại bỏ chúng.

```
1 processed_data = na.omit(processed_data)  
2 dim(processed_data)
```

Thực thi đoạn mã này ta còn lại 405 quan trắc.

- Kiểm tra ngoại lai và cực ngoại lai Đối với bước này, ta chỉ kiểm tra đối với các biến có giá trị là numerics, như vậy ta sẽ khảo sát các biến **like**

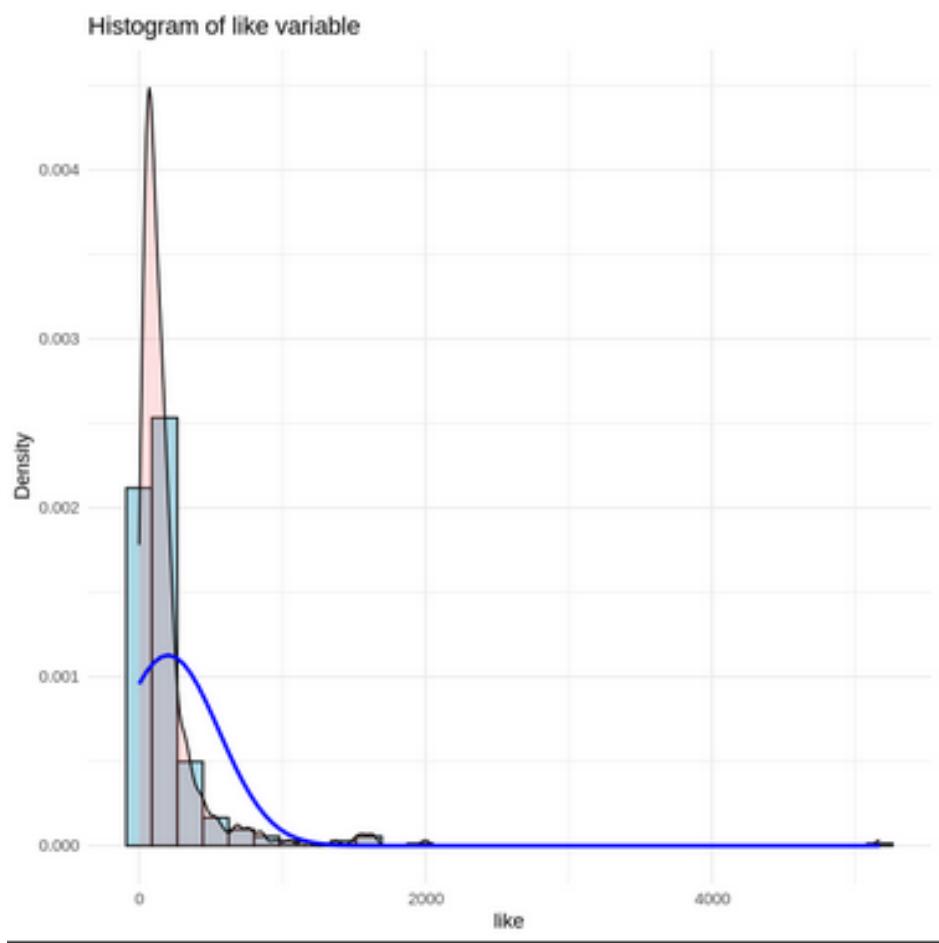
```
1 # Create a box plot  
2 boxplot(processed_data["like"], main="Outliers Analysis  
", col="lightblue")
```



Hình 3.127: Khảo sát outliers

Từ biểu đồ hộp, ta có nhận xét sau đây:Nhìn vào biểu đồ này, chúng ta thấy có rất nhiều điểm ngoại lai và cực ngoại lai ở phía trên, chứng tỏ rằng có một số bài đăng trên facebook có sự tương tác rất mạnh.

Tiếp theo là biểu đồ phân phối của biến **like**



Hình 3.128: Biểu đồ phân phối biến like

Nhận xét: Phân bố có vẻ hơi lệch về phía bên trái so với trung bình.

Bước 3 : Trực quan hóa dữ liệu và rút ra nhận xét.

Ta sẽ dùng R để vẽ ra biểu đồ phân bố của dữ liệu

```

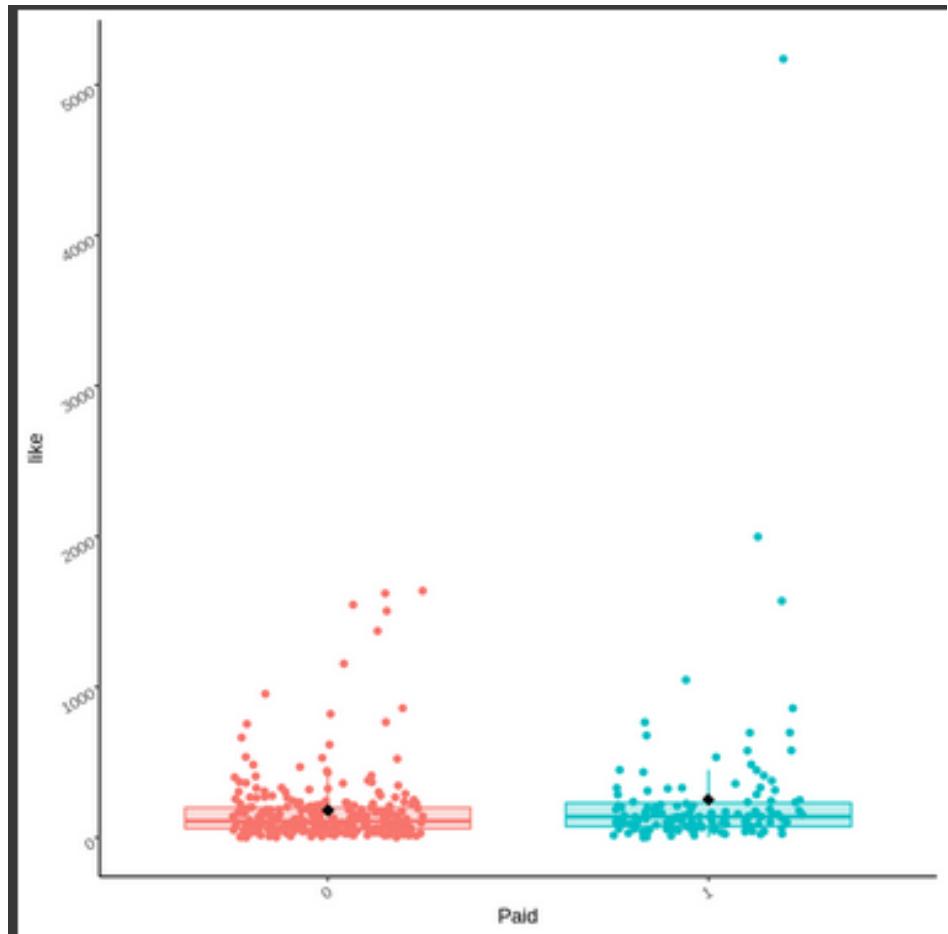
1 # Đưa về kiểu dữ liệu phù hợp
2 processed_data$Paid = factor(processed_data$Paid)
3 processed_data$Category = factor(processed_data$Category)
4
5 # Visualize biến Paid
6 ggplot(processed_data ,aes(x=Paid, y=like, colour=Paid,
+ fill=Paid))+ 
+   geom_jitter(width=0.25)+ 
+   geom_boxplot(alpha=0.25, outlier.alpha=0) + 
+   stat_summary(fun.y=mean, colour="black", geom="point",
+               shape=18, size=3, show.legend = FALSE) +

```

```

11 theme_classic() +
12 theme(legend.position="none")+
13 theme(axis.text = element_text(angle=30, hjust=1, vjust
=1))

```



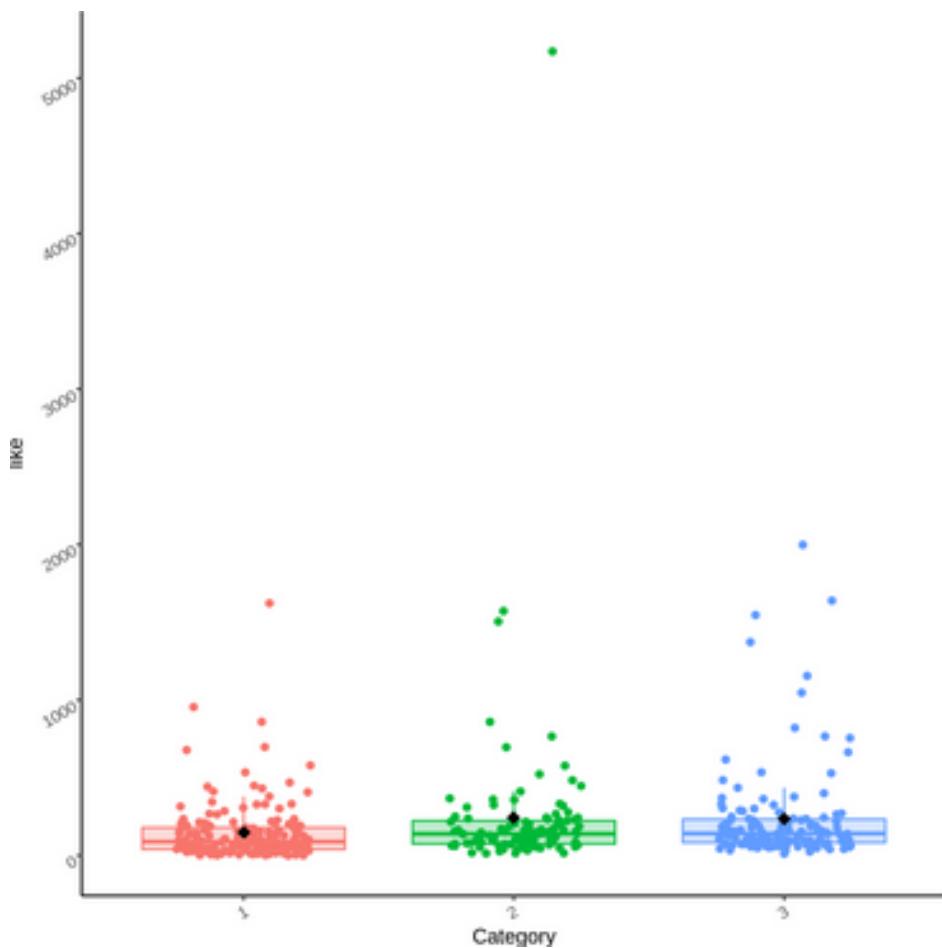
Hình 3.129: Biểu đồ phân bối biến Paid

Từ hai đồ thị trên, ta có một số nhận xét như sau:

- Ở nhóm 0 (Không thuê quảng cáo): Trung vị lớn hơn 0, đa số tập trung trong box; tồn tại ngoại lệ và cực ngoại lệ
- Ở nhóm 1 (Có thuê quảng cáo): Trung vị lớn hơn 0 và lớn hơn nhóm không thuê quảng cáo, chứng tỏ rằng việc chi trả tiền cho quảng cáo sẽ mang lại kết quả tích cực hơn. Số lượng ngoại lai và cực ngoại lai ít hơn nhóm 0 nhưng biến động hơn ở phía trên.
- Việc chi trả tiền thuê quảng cáo về mặt tổng quan cho hiệu quả tích cực hơn so với không thuê quảng cáo, tuy nhiên không phải lúc nào cũng giúp cho bài post tăng sự tương tác (có nhiều điểm gần điểm 0)

Tiếp theo ta sẽ biểu diễn cho biến Category:

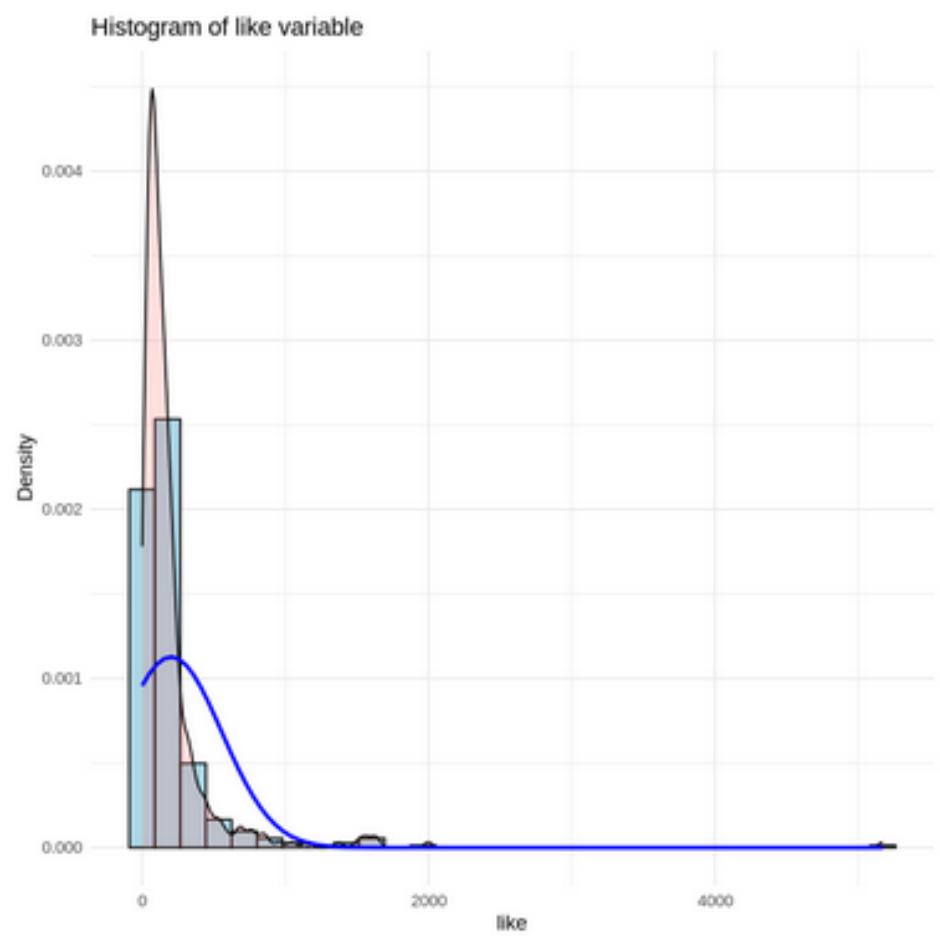
```
1 # Biến Category
2 ggplot(processed_data ,aes(x=Category , y=like , colour=
3   Category ,fill=Category))+ 
4   geom_jitter(width=0.25)+ 
5   geom_boxplot(alpha=0.25 , outlier.alpha=0) + 
6   stat_summary(fun.y=mean , colour="black" , geom="point"
7   ,
8   shape=18 , size=3 , show.legend = FALSE)+ 
9   theme_classic()+
  theme(legend.position="none")+
  theme(axis.text = element_text(angle=30 , hjust=1 ,
  vjust=1))
```



Hình 3.130: Biểu đồ phân bố biến Category

Nhận xét: Về tổng quan ta thấy rằng các nhóm đa số đều nằm trong khoảng box, vẫn tồn

tại các điểm ngoại lai, Tuy nhiên ở chủ đề 2 và 3 cho thấy rằng hiệu quả tích cực hơn nhóm 1 (trung vị cao hơn và biến động ngoại lai rộng hơn) Cuối cùng là biến Like



Hình 3.131: Phân phối biến Like

1. Thống Kê Mô Tả:

- * Giá trị nhỏ nhất: 0.0
- * Phân vị thứ nhất (Q1): 63.0
- * Trung vị (Q2): 118.0
- * Giá trị trung bình: 202.3
- * Phân vị thứ ba (Q3): 210.0
- * Giá trị lớn nhất: 5172.0

2. Hình Dạng Phân Bố:

- * Phân bố có độ lệch phải lớn (lệch dương). Điều này thể hiện qua đuôi dài mở rộng về phía các giá trị lớn hơn.
- * Phần lớn các điểm dữ liệu tập trung ở phía đầu dưới, với đỉnh nhọn gần giá trị nhỏ nhất.

3. Xu Hướng Trung Tâm:

- * Trung vị (118.0) thấp hơn đáng kể so với giá trị trung bình (202.3), đây là một dấu hiệu khác của sự lệch.
- * Các giá trị phân vị thứ nhất và thứ ba cũng cho thấy phần lớn dữ liệu tập trung ở khoảng giá trị thấp.

4. Giá Trị Ngoại Lệ:

- * Có một số giá trị ngoại lệ lớn kéo dài đến 5172.0, xa so với phần lớn dữ liệu. Những ngoại lệ này làm ảnh hưởng đến giá trị trung bình, làm cho nó cao hơn so với trung vị.

5. Biểu Đồ Mật Độ:

- * Đường màu xanh biểu thị ước lượng mật độ của biến "like".
- * Biểu đồ mật độ cũng xác nhận sự lệch, với đỉnh cao ở các giá trị thấp và giảm dần về phía các giá trị cao.

3.3.5. Kiểm định các giả thiết thống kê (ANOVA assumptions)

Nhắc lại các điều kiện để phân tích ANOVA như sau:

1. Các mẫu độc lập
2. Biến phụ thuộc là biến liên tục
3. Các nhóm có phân phối chuẩn hoặc gần chuẩn, đồng nghĩa với việc kiểm định phương sai các nhóm cho kết quả là đồng nhất.

Rõ ràng, theo như phân tích phía trên, bộ dữ liệu chúng ta đã thỏa mãn điều kiện (1) và (2). Để chắc chắn ta sẽ đi kiểm định yêu cầu số (3) bằng cách tiến hành thực hiện các kiểm định sau:

- Shapiro-Wilk test
- leveneTest
- durbinWatsonTest

Đầu tiên ta sẽ xây dựng mô hình tương tác bằng dòng lệnh sau

```
1 # Xây dựng mô hình tương tác
2 int_model = aov(like~Category * Paid, data = processed_data
                  )
3 summary(int_model)
```

Kết quả thu được:

	Df	Sum Sq	Mean Sq	F value	Pr (>F)	
Category	1	584670	584670	4.741	0.0300	*
Paid	1	487829	487829	3.956	0.0474	*
Category : Paid	1	971	971	0.008	0.9293	
Residuals	401	49452484	123323			

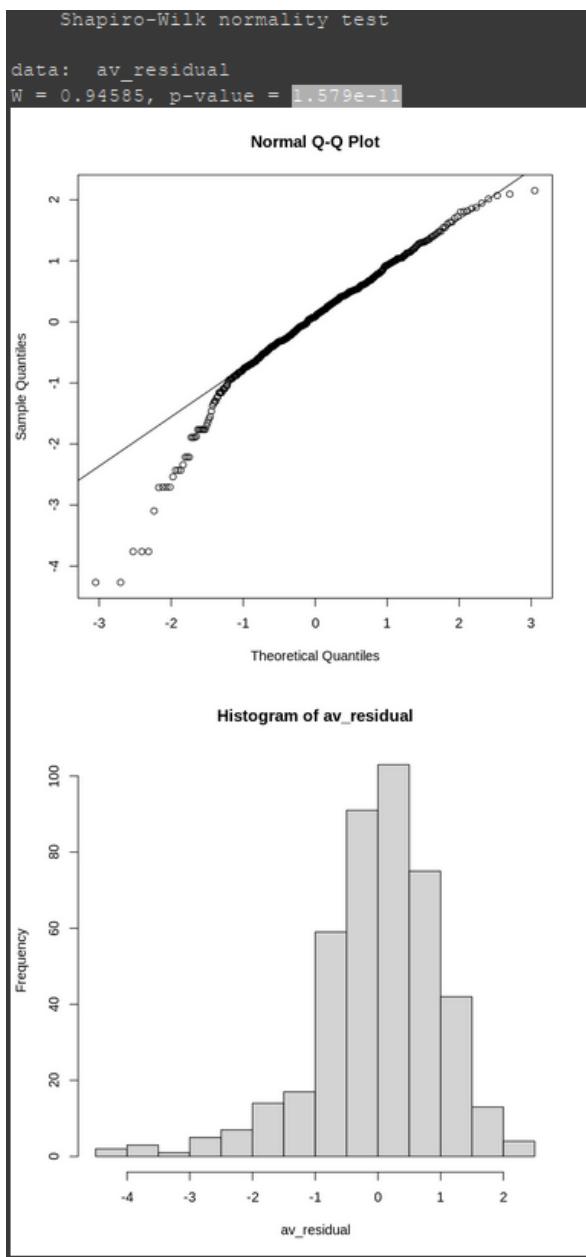
Signif. codes:	0	'***'	0.001	'**'	0.01	'*'
					0.05	'.'
					0.1	'
						1

Với mức ý nghĩa 5%, ta thấy rằng giữa ‘Category’ và ‘Paid’ có mối quan hệ tương tác với nhau dẫn đến tác động hiệu quả của việc dùng thuốc đối với trí nhớ của người sử dụng. Tiếp tục kiểm định các thông số sau:

```

1 # Shapiro-Wilk test
2 av_residual = rstandard(int_model)
3 shapiro.test(av_residual)
4 # Trực quan bằng QQ plot
5 qqnorm(av_residual)
6 qqline(av_residual)
7 hist(av_residual)

```



Hình 3.132: Biểu đồ phần dư

Kết quả như sau:

```
1      Shapiro-Wilk normality test
2
3 data: av_residual
4 W = 0.94585, p-value = 1.579e-11
```

Xét giả định

- H0: Tuân theo phân phối chuẩn
- H1: Không tuân theo phân phối chuẩn

Nhận xét: với độ tin cậy 5% thì với giá trị p-value = 1.579e-11 chúng ta đủ cơ sở bác bỏ H0, vậy sai số có phân phối không chuẩn. Nhìn vào biểu đồ, ta thấy rằng ở phần đuôi kéo dài, nhiều điểm bị kéo lệch ra khỏi đường thẳng, Khả năng các điểm nhiễu chính là các điểm ngoại lệ (outliners).

Bước tiếp theo ta Kiểm định các nhóm có phương sai đồng nhất hay không

```
1 leveneTest(int_model)
```

Kết quả

```
1 A anova: 2 x 3
2 Df       F value Pr(>F)
3           <int>     <dbl>    <dbl>
4 group     5        3.431213    0.004772486
5         399        NA        NA
```

Giả định:

- H0: Các nhóm có phương sai đồng nhất
- H1: Các nhóm không có phương sai đồng nhất

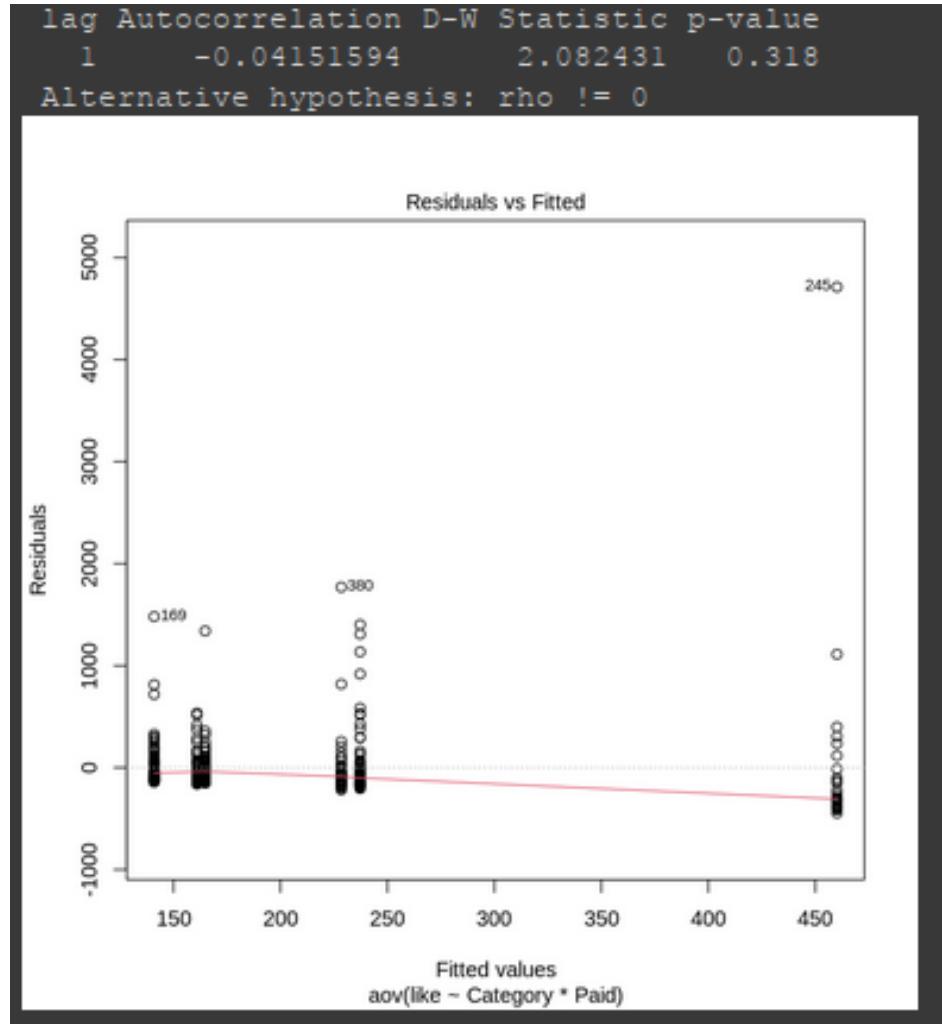
Nhận xét: Với giá trị p-value = 0.004 < 0.05, ta đủ điều kiện bác bỏ H0, vậy các nhóm có phương sai không đồng nhất.

```

1 # Kiểm định tính độc lập của phần dư
2 durbinWatsonTest(int_model)
3 plot(int_model, 1)

```

Kết quả



Hình 3.133: Kết quả kiểm định durbinWatsonTest

Với giả định:

- H₀: Không có sự tương quan (độc lập).
- H₁: H₁: Có sự tương quan (không độc lập).

Thì với giá trị p-value = 0.318 (> 0.05) nên không có sự tương quan. Vậy phần dư độc lập

3.3.6. Phân tích phương sai k nhân tố

Với bước kiểm định levene ta thấy rằng mô hình chúng ta không đảm bảo tính chuẩn, tuy nhiên với mức giá trị $p=0.047$ ta thấy rằng xấp xỉ chuẩn với độ tự tin 0.05. Vì thế, ta vẫn có thể tiếp tục đi phân tích ANOVA. Tiếp theo chung ta sẽ tiến hành đi phân tích phương sai k nhân tố. Việc này gồm các bước sau:

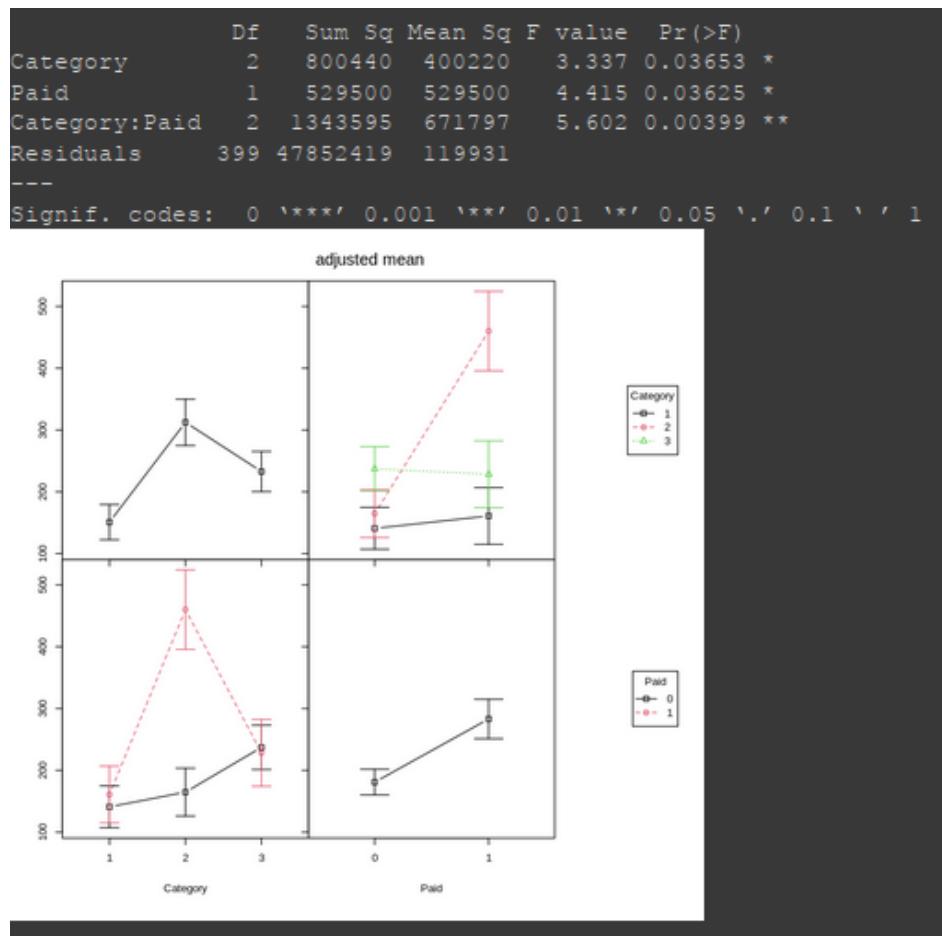
1. Kiểm tra sự tương tác
2. Phân tích ảnh hưởng đơn
 - * Phân tích ảnh hưởng đơn của quảng cáo ở mỗi loại thẻ loại
 - * Phân tích ảnh hưởng đơn của thẻ loại trong việc sử dụng quảng cáo
3. Phân tích ảnh hưởng chính
 - * Phân tích ảnh hưởng chính của Quảng cáo với hiệu quả của bài post
 - * Phân tích ảnh hưởng chính của Thẻ loại với hiệu quả của bài post

Sau đây là các bước chi tiết:

- **Bước 1: Xây dựng mô hình tương tác (interaction model) và kiểm tra tương tác của các biến**

```
1 int_model = aov(like ~ Category * Paid, data = processed_
    data)
2 summary(int_model)
3 plot(interactionMeans(int_model))
```

Kết quả



Hình 3.134: Tương tác giữa các biến trong mô hình

Ta rút ra một số nhận xét dựa trên kết quả như sau:

- Với mức ý nghĩa 5%, Kết quả ANOVA cho thấy các yếu tố "Category" và "Paid" đều có ảnh hưởng đáng kể đến biến phụ thuộc ($p\text{-value} < 0.05$). Tương tác giữa "Category" và "Paid" cũng có ảnh hưởng đáng kể ($p\text{-value} < 0.01$).
- Đối với biểu đồ bên trái:
 - + Đồ thị trên cùng bên trái (Thể hiện mối quan hệ giữa "Category" và giá trị trung bình điều chỉnh): "Category 2" có giá trị trung bình cao nhất, trong khi "Category 1" và "Category 3" có giá trị trung bình thấp hơn.
 - + Đồ thị dưới cùng bên trái (Thể hiện mối quan hệ giữa "Category" và "Paid"):
 - "Category 2" có giá trị trung bình cao nhất khi "Paid- 0, và giảm khi "Paid- 1.
 - Khi Paid = 0, ta thấy rằng có sự tăng trưởng khi Category đi từ 1 đến 3. Ngược lại "Paid- 1" thì kết quả lúc tăng lúc giảm
 - Về tổng quan thì Paid = 1 sẽ cho kết quả tốt hơn, đặc biệt là category 2.
- Đối với biểu đồ bên phải:

- + Đồ thị trên cùng bên phải (Thể hiện mối quan hệ giữa "Paid" và giá trị trung bình điều chỉnh cho các nhóm "Category"): Nhóm category2 cho thấy hiệu quả vượt bậc khi Paid = 1, trong khi 2 nhóm còn lại thì không có sự thay đổi không đáng kể (tăng giảm rất ít).
- + Đồ thị dưới cùng bên phải (Thể hiện mối quan hệ giữa "Paid" và giá trị trung bình điều chỉnh): Tương tác giữa "Category" và "Paid" cho thấy sự thay đổi trong giá trị trung bình giữa các nhóm "Category" phụ thuộc vào trạng thái "Paid".

Ta có kết luận sau đây:

- * Cả "Category" và "Paid" đều ảnh hưởng đáng kể đến giá trị trung bình điều chỉnh của biến phụ thuộc.
- * Tương tác giữa "Category" và "Paid" cho thấy sự thay đổi trong giá trị trung bình giữa các nhóm "Category" phụ thuộc vào trạng thái "Paid".

- **Bước 2: Phân tích ảnh hưởng đơn**

Để phân tích ảnh hưởng đơn ta sẽ sử dụng hàm **testInteractions** để tiến hành phân tích. Sau đây là các bước chi tiết

- * Phân tích ảnh hưởng đơn của quảng cáo ở mỗi loại thẻ loại

```
1 testInteractions(int_model, fixed = "Paid", across =
                    "Category")
```

Kết quả:

Warning message in rbind(deparse.level, ...): "number of columns of result, 8, is not a multiple of vector length 6 of arg 2" A anova: 3 × 8								
	Category1	Category2	SE1	SE2	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
0	-96.24240	-72.4043	49.31306	5.280829e+01	2	482827.1	2.012939	0.134949872
1	-67.43218	231.6888	70.91681	8.402792e+01	2	1750797.0	7.299192	0.001540442
Residuals	NA	NA	399.00000	4.785242e+07	NA	NA	NA	NA

Hình 3.135: Kết quả tương tác giữa Paid và Category

Với các giả định như sau:

- H0: Thuê quảng cáo Không ảnh hưởng đến hiệu quả tương tác bài post
- H1: Thuê quảng cáo Có ảnh hưởng đến hiệu quả của tương tác bài post

Ta rút ra kết luận như sau: Với kết quả phân tích ta có một số nhận xét như sau, với độ tin cậy 5% thì: Thuê quảng cáo có ảnh hưởng đến kết quả của loại tương tác bài post ở các thẻ loại

- * Phân tích ảnh hưởng đơn của thẻ loại ở việc thuê quảng cáo

```

1 testInteractions(int_model, fixed = "Category",
                  across = "Paid")

```

Warning message in rbind(deparse.level, ...): "number of columns of result, 6, is not a multiple of vector length 5 of arg 2" Anova: 4 x 6						
	Value	SE	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	-20.020551	56.97587	1	14808.164	0.12347249	1.000000000
2	-295.303448	75.06453	1	1856087.831	15.47631351	0.000295576
3	8.789667	64.92090	1	2198.401	0.01833056	1.000000000
Residuals	NA	399.00000	47852419	NA	NA	NA

Hình 3.136: Ảnh hưởng đơn giữa thẻ loại ở việc thuê quảng cáo

Tương tự như ở phía trên, ta có các giả định như sau:

- H0: Thẻ loại không có sự tương tác trong việc thuê quảng cáo
- H1: Thẻ loại có sự tương tác trong việc thuê quảng cáo

Ta rút ra kết luận như sau: Với kết quả phân tích ta có một số nhận xét như sau, với độ tin cậy 5% thì: Chỉ có thẻ loại thứ 2 thẻ hiện rõ sự tương tác với **Paid**, các trường hợp còn lại là không thẻ hiện sự tương tác.

* Phân tích ảnh hưởng đơn giữa các nhóm thẻ loại ứng với việc thuê quảng cáo hay không thuê quảng cáo

Việc phân tích sự tương tác của các nhóm trong cùng một điều kiện nhất định cũng có ý nghĩa rất quan trọng trong thống kê, từ đó sẽ hiểu rõ hơn về từng tác dụng của từng loại và từng nhóm

```

1 options(contrasts = c(unordered="contr.sum", ordered
                        ="contr.poly"))
2 one_vs_two = list(Category = c(1, -1, 0))
3 one_vs_three = list(Category = c(1, 0, -1))
4 two_vs_three = list(Category = c(0, 1, -1))

```

Dầu tiên, ta sẽ đi phân tích ảnh hưởng của nhóm Category1 và Category2

```

1 testInteractions(int_model, custom = c(one_vs_two),
                   fixed = "Paid", adjustment = "bonferroni")

```

Kết quả:

Warning message in rbind(deparse.level, ...): "number of columns of result, 6, is not a multiple of vector length 5 of arg 2" Anova: 3 x 6						
	Value	SE	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
0 : Category1	-23.8381	51.39392	1	25801.84	0.2151393	1.00000000000
1 : Category1	-299.1210	78.99113	1	1719761.36	14.3396048	0.0003521958
Residuals	NA	399.00000	47852419	NA	NA	NA

Hình 3.137: Tương tác giữa nhóm 1 và 2 ở mỗi liều lượng

Ta có giả định như sau:

- H0: Không có sự khác nhau giữa nhóm 1 và nhóm 2
- H1: Có sự khác nhau giữa nhóm 1 và nhóm 2

Ta rút ra kết luận như sau: Với kết quả phân tích ta có một số nhận xét như sau, với độ tin cậy 5% thì: Có sự tương tác có ý nghĩa thống kê ở việc thuê quảng cáo giữa nhóm 1 và nhóm 2.

Tiếp theo là nhóm Category1 và Category3

```
1 testInteractions(int_model, custom = c(one_vs_three)
, fixed = "Paid", adjustment = "bonferroni")
```

Kết quả:

Warning message in rbind(deparse.level, ...): "number of columns of result, 6, is not a multiple of vector length 5 of arg 2" Anova: 3 x 6						
	Value	SE	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
0 : Category1	-96.24240	49.31306	1	456814.5	3.8089819	0.1033554
1 : Category1	-67.43218	70.91681	1	108434.4	0.9041407	0.6844990
Residuals	NA	399.00000	47852419	NA	NA	NA

Hình 3.138: Tương tác giữa nhóm 1 và 3

Với các giả định tương tự với nhóm 1 và 2, ta có kết luận như sau:

- Có sự tương tác có ý nghĩa thống kê ở việc thuê quảng cáo giữa nhóm 1 và nhóm 3
- Có sự tương tác có ý nghĩa thống kê ở việc không thuê quảng cáo giữa nhóm 1 và nhóm 3

Cuối cùng là giữa nhóm 2 và 3

```
1 testInteractions(int_model, custom = c(two_vs_three)
, fixed = "Paid", adjustment = "bonferroni")
```

Kết quả:

Warning message in rbind(deparse.level, ...): "number of columns of result, 6, is not a multiple of vector length 5 of arg 2"						
A anova: 3 x 6						
	Value	SE	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
0 : Category1	-72.4043	52.80829	1	225452.8	1.879856	0.3422424
1 : Category1	231.6888	84.02792	1	911788.2	7.602614	0.0121922
Residuals	NA	399.00000	47852419	NA	NA	NA

Hình 3.139: Tương tác giữa nhóm 2 và 3

Với các giả định tương tự với nhóm 1 và 2, ta có kết luận như sau: Với độ tin cậy 5% thì

- Có sự tương tác có ý nghĩa thống kê ở việc thuê quảng cáo giữa nhóm 2 và nhóm 3
- Có sự tương tác có ý nghĩa thống kê ở việc không thuê quảng cáo giữa nhóm 2 và nhóm 3

Từ việc phân tích trên, ta có kết luận như sau: Việc thuê quảng cáo chỉ cho kết quả tốt hơn ở nhóm 1 và 2. Các nhóm còn lại việc chi tiền cho quảng cáo và không chi tiền đều cho kết quả như nhau, Khả năng cao mức ảnh hưởng này sẽ giải thích bởi một yếu tố khác mà ta chưa khảo sát (như content chẳng hạn).

- Phân tích ảnh hưởng đơn giữa quảng cáo ứng với các nhóm thể loại

```

1 options(contrasts = c(unordered="contr.sum", ordered=
2   "contr.poly"))
3 no_vs_yes = list(Paid = c(1, -1))
4 testInteractions(int_model, custom = c(no_vs_yes),
5   fixed = "Category", adjustment = "bonferroni")

```

Kết quả:

Warning message in rbind(deparse.level, ...): "number of columns of result, 6, is not a multiple of vector length 5 of arg 2"						
A anova: 4 x 6						
	Value	SE	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1 : Paid1	-20.020551	56.97587	1	14808.164	0.12347249	1.000000000
2 : Paid1	-295.303448	75.06453	1	1856087.831	15.47631351	0.000295576
3 : Paid1	8.789667	64.92090	1	2198.401	0.01833056	1.000000000
Residuals	NA	399.00000	47852419	NA	NA	NA

Hình 3.140: Ảnh hưởng đơn giữa quảng cáo ứng với các nhóm thể loại

Với giả định sau:

- * H0: Không có sự khác nhau trong tương tác hiệu quả giữa việc thuê quảng cáo và không thuê quảng cáo

- * H1: Có sự khác nhau trong tương tác hiệu quả giữa việc thuê quảng cáo và không thuê quảng cáo

Với độ tin cậy 5% thì: Ở thẻ loại 1 và 3, chỉ ra không có sự khác nhau hiệu quả giữa việc thuê và không thuê quảng cáo, trong khi đó thẻ loại 2 cho thấy sự khác nhau này về mặt thống kê.

- **Bước 3: Phân tích ảnh hưởng chính**

Ở bước này ta sẽ thực hiện 2 phân tích:

- * Phân tích ảnh hưởng chính của Paid với hiệu quả của bài post thông qua số lượt like
- * Phân tích ảnh hưởng chính của Category với hiệu quả của bài post thông qua lượt like

Với mỗi bước, ta sẽ thực hiện các công việc sau:

- * Xây dựng mô hình
- * Kiểm định các giả thiết của mô hình
- * Kiểm định trung bình của các nhóm
- * Nhận xét

Sau đây là các bước phân tích cụ thể:

- * **Bước 3.1: Phân tích ảnh hưởng chính của Paid với hiệu quả của bài post thông qua số lượt like**

```

1 paid_model = aov(like ~ Paid, data = processed_data)
2 summary(paid_model)

```

Kết quả:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
1 Paid	1	439911	439911	3.54	0.0606 .
2 Residuals	403	50086044	124283		
3 ---					
4 Signif. codes:	0 ***	0.001 **	0.01 *	0.05 .	
5	0.1 ' '	1			

Nhận xét: Với mức ý nghĩa 0.05, ta thấy rằng Paid không có ý nghĩa trong việc giải thích mô hình, tuy nhiên chúng ta vẫn đi kiểm định các giả thiết phía sau (vì với mức ý nghĩa 0.06 cũng khá là gần với 0.05)

Tiếp theo tiến hành kiểm định các giả thuyết

```

1 # Shapiro-Wilk test
2 av_residual = rstandard(paid_model)
3 shapiro.test(av_residual)
4

```

```

5 # Trục quan bằng QQ plot
6 qqnorm(av_residual)
7 qqline(av_residual)
8 hist(av_residual)

```

Kết quả: Với giả định:

- H0: Phần dư Tuân theo phân phối chuẩn.
- H1: Phần dư Không tuân theo phân phối chuẩn.

Nhận xét: với độ tin cậy 5% thì với giá trị p-value = 2.2e-16 chúng ta đủ cơ sở bác bỏ H0, vậy sai số có phân phối không chuẩn. Nhìn vào biểu đồ, ta thấy rằng ở phần đuôi kéo dài, có nhiều điểm bị kéo lệch ra khỏi đường thẳng đặc biệt là đuôi phía trên → Khả năng các điểm nhiễu chính là các điểm ngoại lệ (outliners), biểu đồ lệch chuẩn

```

1 # Kiểm định các nhóm có phương sai đồng nhất hay không
2 leveneTest(paid_model)

```

Kết quả:

```

1          A anova: 2 x 3
2 Df      F value Pr(>F)
3 <int>    <dbl>   <dbl>
4 group     1        2.460602    0.1175189
5           403       NA        NA

```

Với giả định:

- Các nhóm có phương sai đồng nhất.
- Các nhóm không có phương sai đồng nhất.

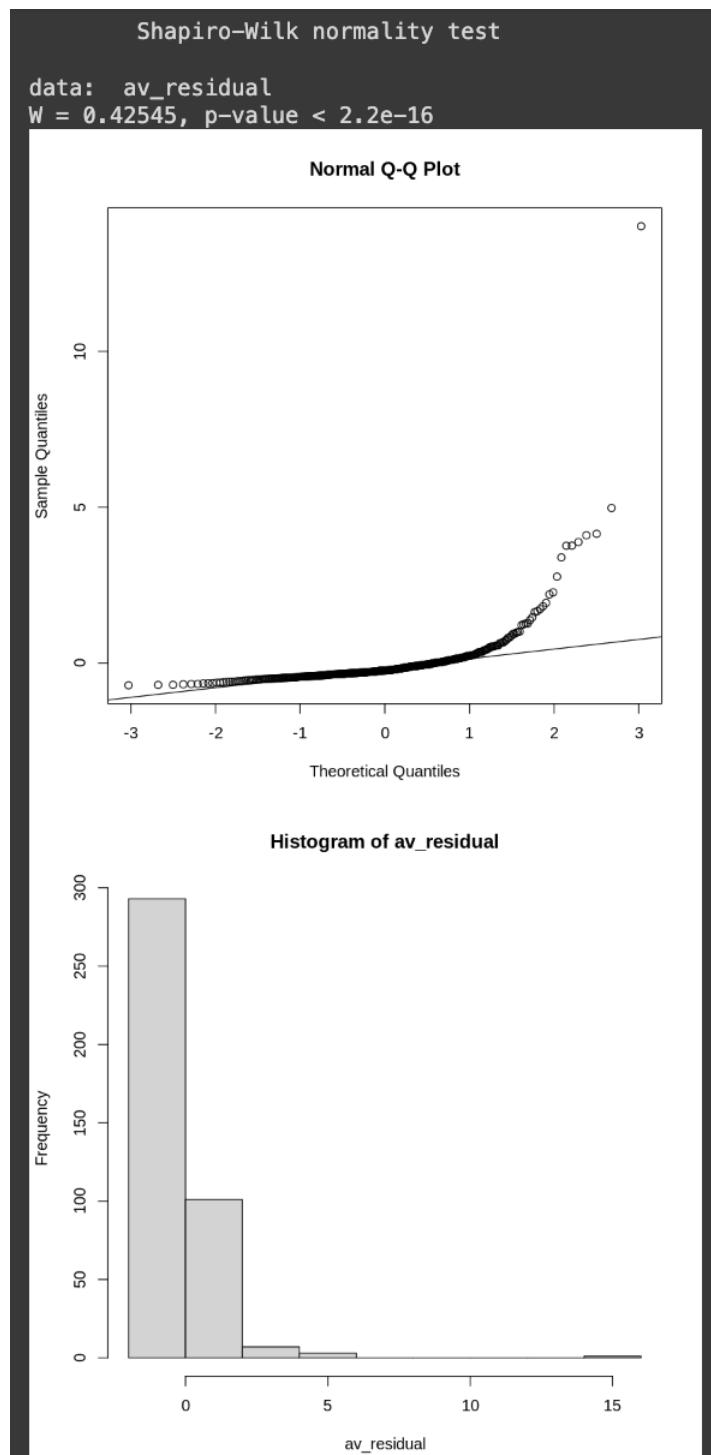
Nhận xét: Với giá trị p-value = 0.1175189 > 0.05, ta không đủ điều kiện bác bỏ H0, vậy các nhóm có phương sai đồng nhất.

```

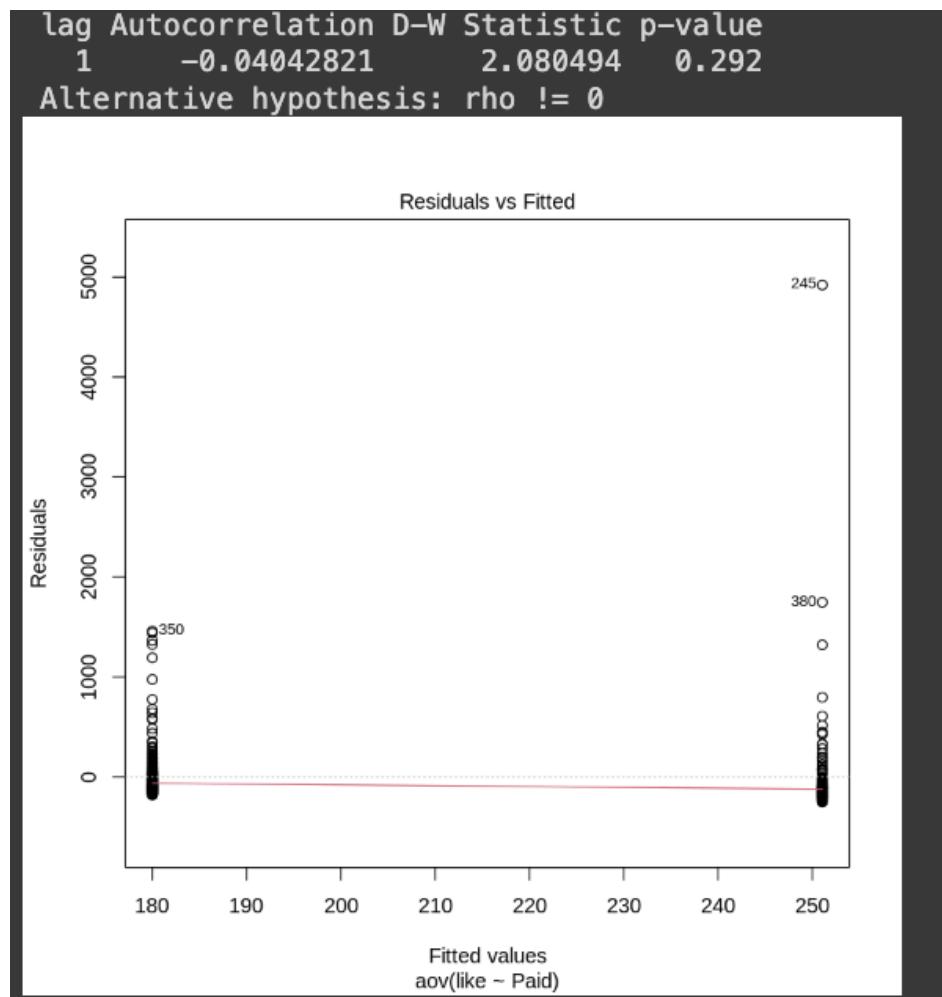
1 # Kiểm định tính độc lập của phần dư
2 durbinWatsonTest(paid_model)
3 plot(paid_model, 1)

```

Kết quả:



Hình 3.141: Kết quả kiểm định Shapiro-Wilk test



Hình 3.142: Kiểm định durbinWatsonTest

Với giả định:

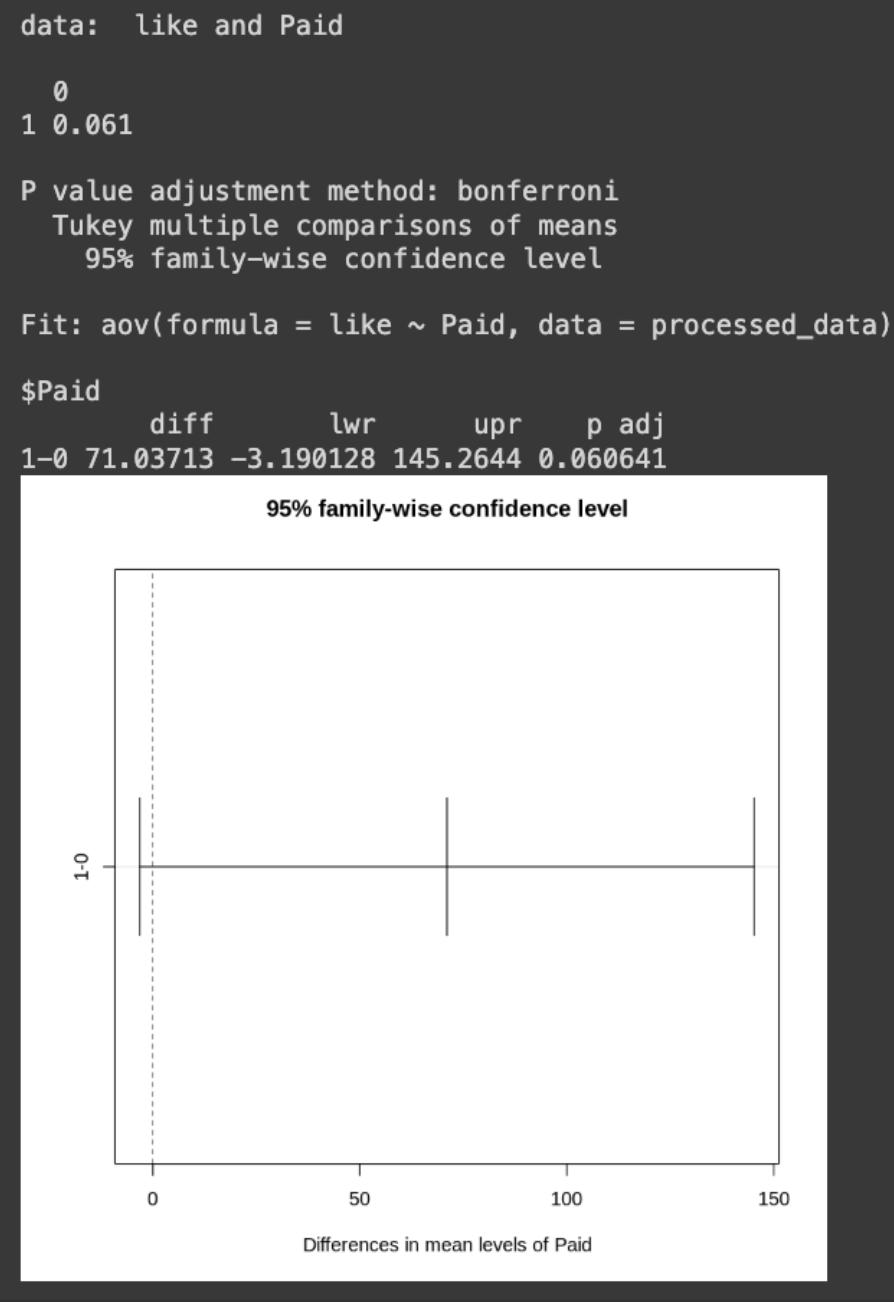
- * H0: Không có sự tương quan (độc lập)
- * H1: Có sự tương quan (không độc lập)

Nhận xét: Với giá trị p-value = 0.292 nên không có sự tương quan.

Mặc dù với điều kiện phương sai giữa các nhóm không đồng nhất nên sẽ không tiến hành phân tích ANOVA được, tuy nhiên về mặc trực quan hóa dữ liệu, ta thấy rằng đồ thị phân bố dạng gần chuẩn, nên ta sẽ tiếp tục đi phân tích các yếu tố ANOVA.

```
1 with(processed_data, pairwise.t.test(like, Paid, p.adj  
= "bonferroni"))  
2 TukeyHSD(aov(like~Paid, data=processed_data), conf.  
level = 0.95)  
3 plot(TukeyHSD(aov(like~Paid, data=processed_data), conf  
.level = 0.95))
```

Kết quả:



Hình 3.143: Kiểm định Tukey's

Với giả định:

- * H₀: Các giá trị trung bình giữa các cặp bằng nhau
- * H₁: Các giá trị trung bình giữa các cặp không bằng nhau

Nhìn vào kết quả ta có:

- * Nhìn vào kết quả ta có: p-value=0.060641 có giá trị lớn hơn 0.05 (độ tin cậy 95%) nên ta không có cơ sở để bác bỏ H₀. Vậy rõ ràng giữa các nhóm này có giá trị trung bình

là như nhau. Nghĩa là các nhóm có quảng cáo hay không quảng cáo thì độ hiệu quả là như nhau thông qua số lượng lượt like.

- * Nhìn vào kết quả và hình vẽ ta cũng thấy ngay giữa nhóm có mức độ hiệu quả trung bình như nhau (đồ thị cắt điểm 0)

Vì chỉ có 2 nhóm nên ta sẽ không tiến hành phân tích tương tác nhóm. Như vậy, ta có kết luận như sau: Việc chi tiền cho quảng cáo hay không cũng chỉ cho cùng một kết quả như nhau (do trung bình như nhau).

- **Bước 3.2: Phân tích ảnh hưởng chính của Category với hiệu quả của bài post thông qua lượt like**

```

1 category_model = aov(like~Category, data = processed_
  data)
2 summary(category_model)

```

Kết quả:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Category	2	800440	400220	3.236	0.0404 *
Residuals	402	49725514	123695		

Signif. codes:	0	'***'	0.001	'**'	0.01 '*' 0.05 '.'
	0.1	' '	1		

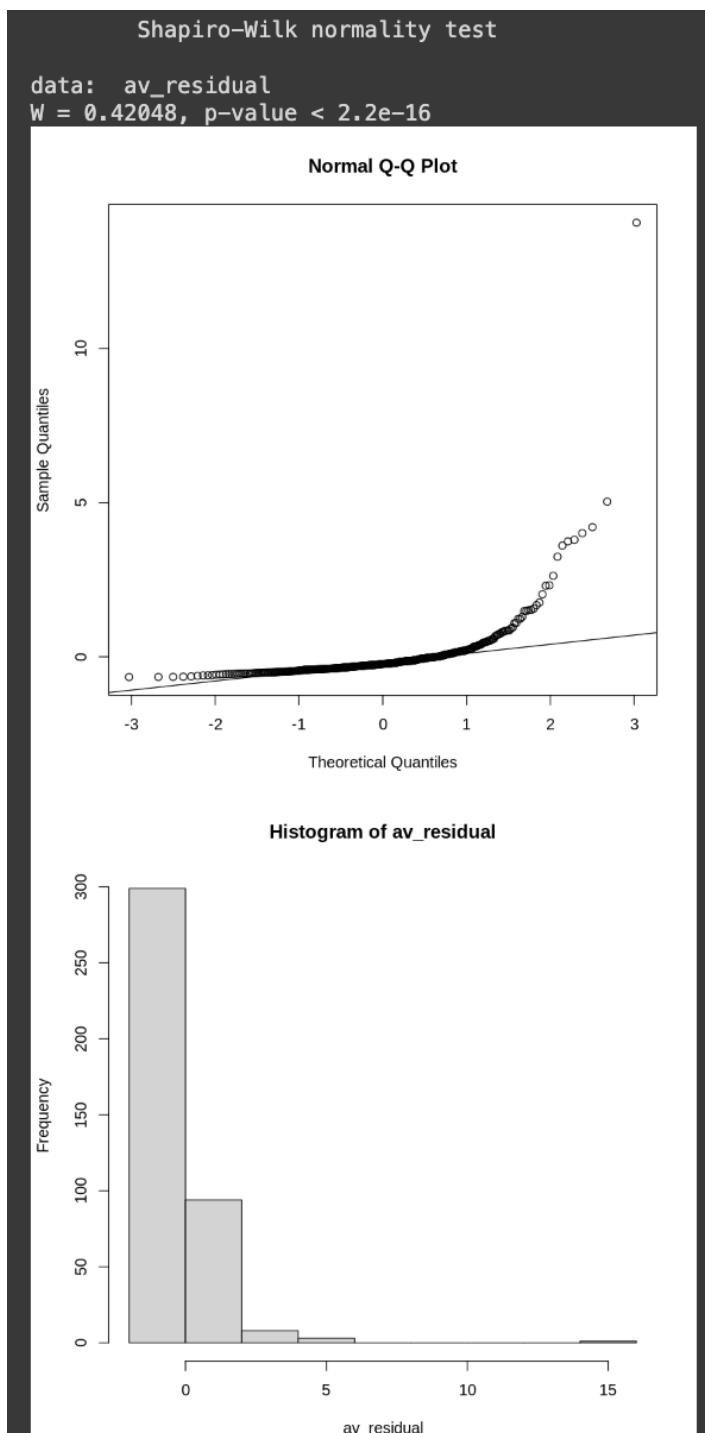
Nhận xét: Với mức ý nghĩa 0.05, ta thấy rằng Category có ý nghĩa trong việc giải thích mô hình

```

1 # Shapiro-Wilk test
2 av_residual = rstandard(category_model)
3 shapiro.test(av_residual)
4
5 # Trực quan bằng QQ plot
6 qqnorm(av_residual)
7 qqline(av_residual)
8 hist(av_residual)

```

Kết quả:



Hình 3.144: Kết quả Shapiro-Wilk test và đồ thị phân phối

Với các giả định:

- * H0: Phản dư tuân theo phân phối chuẩn.
- * H1: Phản dư không tuân theo phân phối chuẩn

Nhận xét: với độ tin cậy 5% thì với giá trị p-value = 2.2e-16 chúng ta đủ cơ sở bác bỏ H0,

vậy sai số có phân phối không chuẩn. Nhìn vào biểu đồ, ta thấy rằng ở phần đuôi kéo dài, có nhiều điểm bị kéo lệch ra khỏi đường thẳng đặc biệt là đuôi phía trên, Khả năng các điểm nhiễu chính là các điểm ngoại lệ (outliers), biểu đồ lệch chuẩn

```
1 # Kiểm định các nhóm có phương sai đồng nhất hay không
2 leveneTest(category_model)
```

Kết quả:

```
1 A anova: 2 x 3
2 Df      F value Pr(>F)
3 <int>    <dbl>   <dbl>
4 group     2        1.191298    0.3048969
5          402       NA       NA
```

Với các giả định:

- * H0: Các nhóm có phương sai đồng nhất
- * H1: Các nhóm không có phương sai đồng nhất

Nhận xét: Với giá trị p-value = 0.3048969 > 0.05, ta không đủ điều kiện bác bỏ H0, vậy các nhóm có phương sai đồng nhất.

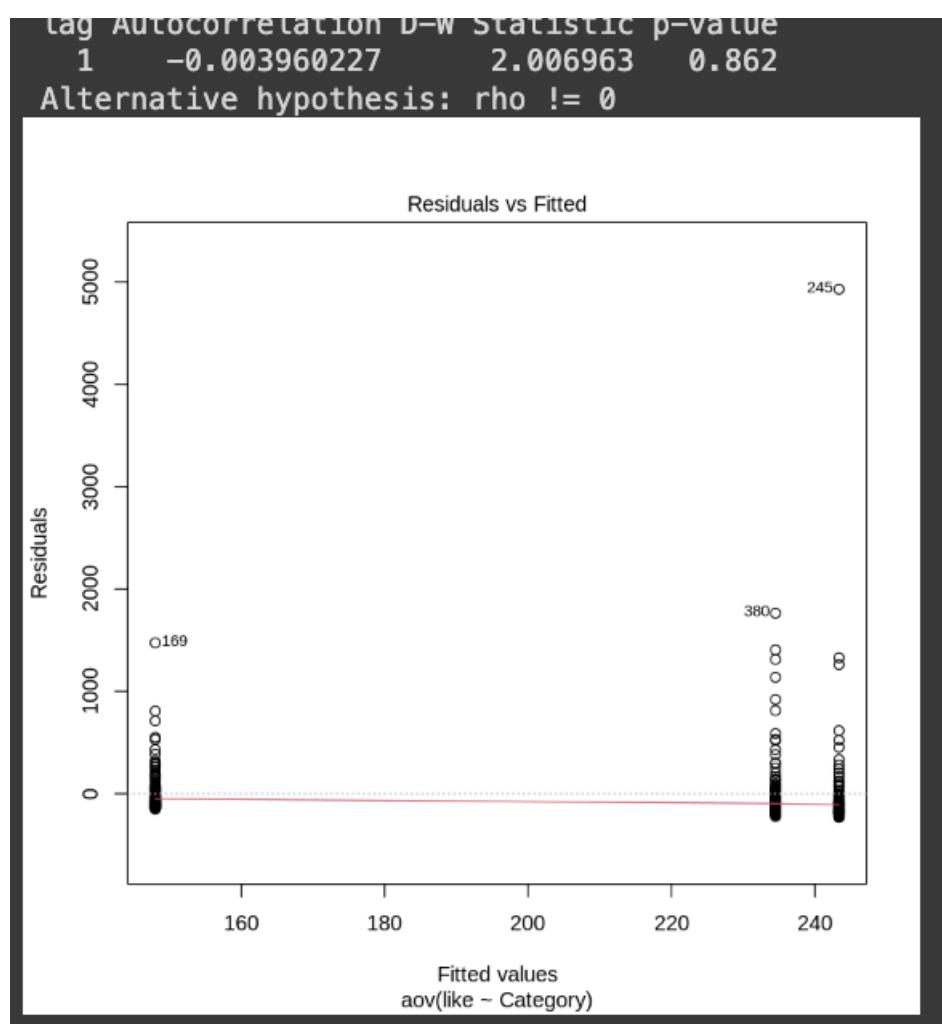
```
1 # Kiểm định tính độc lập của phần dữ
2 durbinWatsonTest(category_model)
3 plot(category_model, 1)
```

Kết quả: Với các giả định:

- * H0: Không có sự tương quan (độc lập)
- * H1: Có sự tương quan (không độc lập)

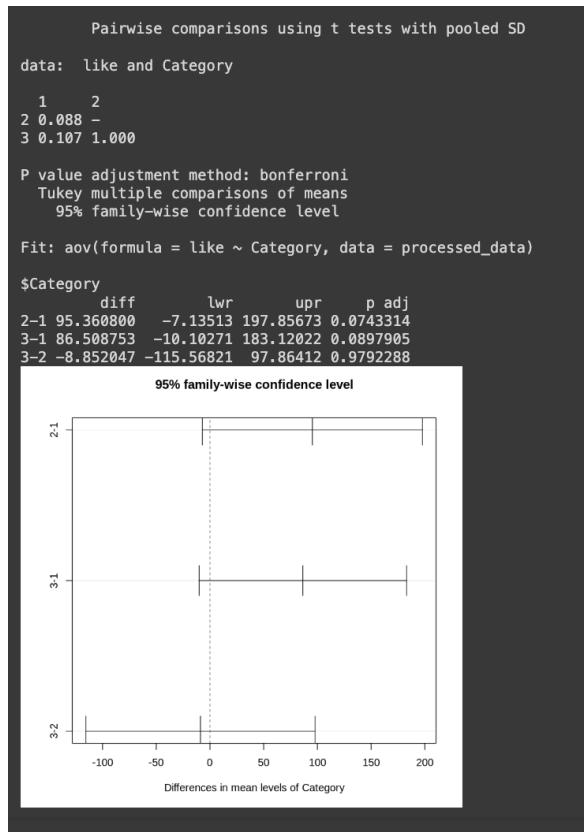
Nhận xét: Nhận xét: Với giá trị p-value = 0.87 nên không có sự tương quan. Mặc dù với điều kiện phương sai giữa các nhóm không đồng nhất nên sẽ không tiến hành phân tích ANOVA được, tuy nhiên về mặc trực quan hóa dữ liệu, ta thấy rằng đồ thị phân bố dạng gần chuẩn, nên ta sẽ tiếp tục đi phân tích các yếu tố ANOVA.

```
1 # Kiểm định độ hiệu quả trung bình giữa các nhóm
2 category
3 with(processed_data, pairwise.t.test(like, Category, p.
4   adj = "bonferroni"))
5 TukeyHSD(aov(like ~ Category, data=processed_data), conf.
6   level = 0.95)
7 plot(TukeyHSD(aov(like ~ Category, data=processed_data),
8   conf.level = 0.95))
```



Hình 3.145: Kiểm định durbinWatsonTest

Kết quả:



Hình 3.146: Kết quả kiểm định giá trị trung bình

Với các giả thuyết:

- * H0: Các giá trị trung bình giữa các cặp bằng nhau
- * H1: Các giá trị trung bình giữa các cặp không bằng nhau

Nhìn vào kết quả ta có:

- * Nhìn vào kết quả ta có: Tất cả các nhóm đều có p-value > 0.05 nên các mức độ trung bình giữa các cặp là như nhau.
- * Nhìn vào kết quả và hình vẽ ta cũng thấy ngay giữa các cặp có mức độ hiệu quả trung bình như nhau (đồ thị cắt điểm 0)

3.3.7. Xây dựng và kiểm định mô hình công (Additive model)

```
1 # Xây dựng mô hình công
2 add_model = aov(like~Category + Paid, processed_data)
3 summary(add_model)
```

Kết quả

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Category	1	838265	838265	8.206	0.00435	**
Paid	1	673770	673770	6.596	0.01051	*
Residuals	495	50564590	102151			

Signif. codes:	0	'***'	0.001	'**'	0.01	'*'
					0.05	'. '
					0.1	'
						1

Nhận xét: Với p-value=5%, các biến đều có ý nghĩa trong giải thích mô hình. Ta tiến hành kiểm định Shapiro và Breusch-Pagan

```

1 # Shapiro-Wilk test
2 av_residual = rstandard(add_model)
3 shapiro.test(av_residual)
4
5 # Trực quan bằng QQ plot
6 qqnorm(av_residual)
7 qqline(av_residual)
8 hist(av_residual)

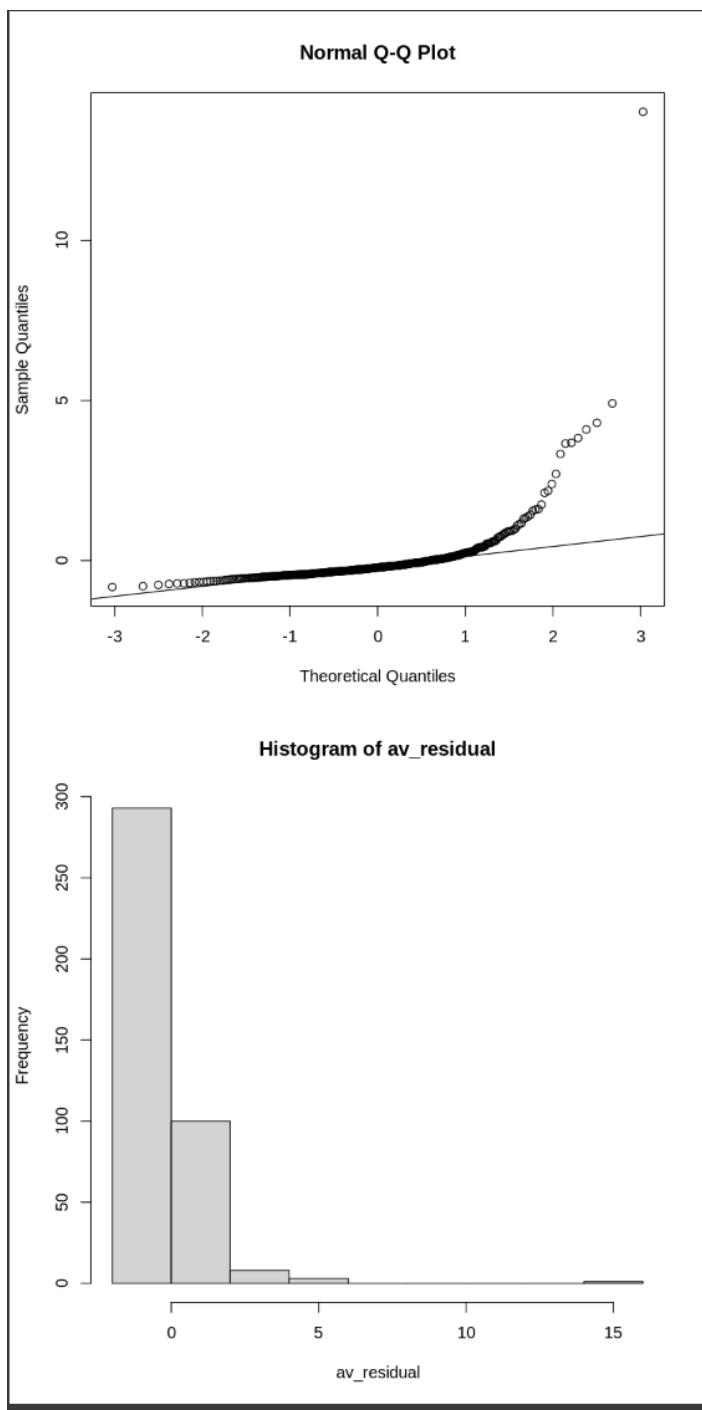
```

Kết quả:

```

1 Shapiro-Wilk normality test
2
3 data: av_residual
4 W = 0.43469, p-value < 2.2e-16

```



Hình 3.147: Shapiro test và biểu đồ chuẩn của phần dư

Với các giả định:

- Phần dư H0: Tuân theo phân phối chuẩn
- H1: Phần dư Không tuân theo phân phối chuẩn

Nhận xét: với độ tin cậy 5% thì với giá trị p-value = 2.2e-16 chúng ta đủ cơ sở bác bỏ H0, vậy

sai số có phân phối không chuẩn. Nhìn vào biểu đồ, ta thấy rằng ở phần đuôi kéo dài, có nhiều điểm bị kéo lệch ra khỏi đường thẳng đặc biệt là đuôi phía trên \rightarrow Khả năng các điểm nhiễu chính là các điểm ngoại lệ (outliners), biểu đồ lệch chuẩn

```
1 # Kiểm định tính độc lập của phần dư  
2 durbinWatsonTest(add_model)  
3 plot(add_model, 1)
```

Kết quả:

```
1 lag Autocorrelation D-W Statistic p-value  
2     1      -0.03610975      2.07151    0.428  
3 Alternative hypothesis: rho != 0
```

Với các giả định:

- H0: Không có sự tương quan (độc lập)
- H1: Có sự tương quan (không độc lập)

Nhận xét: Với giá trị p-value = 0.412 nên không có sự tương quan.

```
1 # Kiểm định Breusch-Pagan  
2 bptest(add_model)
```

Kết quả:

```
1 studentized Breusch-Pagan test  
2  
3 data: add_model  
4 BP = 5.3467, df = 3, p-value = 0.1481
```

Với các giả định:

- H0: phuong sai không đổi
- H1: phuong sai thay doi

Nhận xét: Với p-value=0.148 > 0.05 thì ta không đủ điều kiện bác bỏ H0. Vậy phuong sai của mô hình không thay đổi.

Kết luận: Giữa "Paid" và "Category" có sự tương tác với nhau tác động đến hiệu quả của bài post thông qua số lượt like. Đặc biệt là nhóm category2 nếu dùng quảng cáo sêc cho kết quả tích cực. Trong trường hợp ngược lại thì không nên thuê quảng cáo vì không có sự khác biệt giữa trước và sau thuê.

3.3.8. Cải tiến mô hình

Như chúng ta đã biết, trong quá trình xử lý dữ liệu, ta thấy có một các điểm cực ngoại lai, khả năng cao sẽ ảnh hưởng đến chất lượng mô hình. Vì vậy chúng ta sẽ tiến hành loại bỏ các điểm này.

Đầu tiên ta sẽ tiến hành khảo sát các điểm ngoại lai bằng lệnh sau

```
1 # Khảo sát ngoại lai theo biến diff
2 like_data = processed_data["like"]
3 outliers_index = list()
4 extreme_outliers_index = list()
5
6 for (i in 1:ncol(like_data)) {
7     # Tính toán Q1, Q3 và IQR
8     Q1 = quantile(like_data[, i], 0.25, na.rm = TRUE)
9     Q3 = quantile(like_data[, i], 0.75, na.rm = TRUE)
10    IQR = Q3 - Q1
11
12    # Xác định ngoại lai
13    outliers_index_i = like_data[, i] < (Q1 - 1.5 * IQR) |
14        like_data[, i] > (Q3 + 1.5 * IQR)
15    # outliers_i = like_data[like_data[, i] < (Q1 - 1.5 * IQR)
16        # | like_data[, i] > (Q3 + 1.5 * IQR), i]
17
18    # Lưu trữ ngoại lai
19    field_name = names(like_data)[i]
20    outliers_index[[field_name]] = which(outliers_index_i)
21
22    # Xác định cực ngoại lai
23    extreme_outliers_index_i = like_data[, i] < (Q1 - 3 * IQR)
24        # | like_data[, i] > (Q3 + 3 * IQR)
25    extreme_outliers_index[[field_name]] = which(extreme_
26        outliers_index_i)
27}
```

```

28 print(paste("Số cực ngoại lai:", length(extreme_outliers_
29           index[[names(like_data)[i]]])))
30
31 # Tìm tổng số quan trắc ngoại lai và cực ngoại lai thực sự
32 outliers = c()
33 extreme_outliners = c()
34 for (i in 1:ncol(like_data)){
35   outliers = c(outliers, outliers_index[[names(like_data)
36             [i]]])
37   extreme_outliners = c(extreme_outliners, extreme_
38             outliers_index[[names(like_data)[i]]]))
39 }
40
41 outliers = unique(outliers)
42 extreme_outliners = unique(extreme_outliners)
43 print(paste("Tổng số ngoại lai:", length(outliers)))
44 print(paste("Tổng số cực ngoại lai:", length(extreme_
45           outliners)))

```

Kết quả:

```
1 [1] "Biến: like"
2 [1] "Số ngoại lai: 40"
3 [1] "Số cực ngoại lai: 22"
4 [1] "Tổng số ngoại lai: 40"
5 [1] "Tổng số cực ngoại lai: 22"
```

Như vậy, tổng số ngoại lai và cực ngoại là là 62 samples (chiếm khoảng 12%). Ta tiến hành loại bỏ các điểm này

```
1 # Loại bỏ các điểm ngoại lai và cực ngoại lai
2 rm_outliner_data = processed_data[-extreme_outliners,]
3 rm_outliner_data = rm_outliner_data[-outliers,]
4
5 # Kiểm tra lại số lượng dữ liệu
6 dim(rm_outliner_data)
7 str(rm_outliner_data)
```

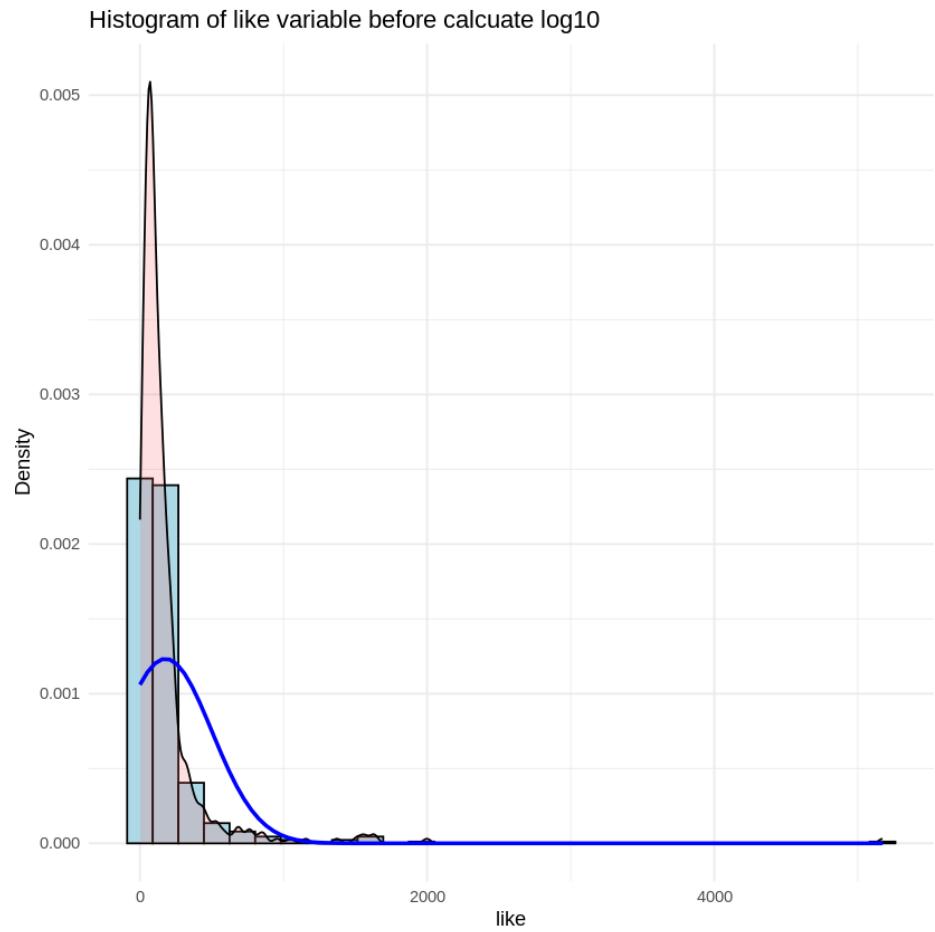
Kết quả

```
1 4383
2 'data.frame':   438 obs. of  3 variables:
3   $ Category: Factor w/ 3 levels "1","2","3": 2 2 3 2 3 3 2
4     3 2 2 ...
5   $ Paid      : Factor w/ 2 levels "0","1": 1 1 1 1 2 2 1 1 1
6     1 ...
7   $ like      : int  79 130 66 152 249 325 161 113 233 88 ...
```

Như vậy, sau khi loại bỏ các điểm ngoại lai ta thu còn lại 438 samples. Ta tiến hành trực quan hóa đồ thị của dữ liệu

```
1 # Biến phụ thuộc Diff
2 ggplot(rm_outliner_islander, aes(x = Diff)) +
3   geom_histogram(aes(y = ..density..), bins = 30, color = "black",
4                 fill = "lightblue") +
5   geom_density(alpha = 0.2, fill = "#FF6666") +
6   stat_function(fun = dnorm, args = list(mean = mean(rm_
7     outliner_islander$Diff, na.rm = TRUE), sd = sd(rm_
8     outliner_islander$Diff, na.rm = TRUE)),
9                 color = "blue", size = 1) +
10  theme_minimal() +
11  labs(title = "Histogram of Diff variable", x = "Diff", y =
12       = "Density")
13 summary(rm_outliner_islander$Diff)
```

Kết quả:



Hình 3.148: Biểu đồ trước khi loại bỏ ngoại lai

Nhận xét: Sau khi loại bỏ các điểm ngoại lai và cực ngoại lai, ta thu được đồ thị gần chuẩn và có hình dáng tốt hơn trước khi loại.

Tiếp theo chúng ta sẽ tiến hành xây dựng mô hình tương tác và kiểm định các giả thuyết

```

1 # Shapiro-Wilk test
2 int_model = aov(like ~ Category * Paid, rm_outliner_data)
3 summary(int_model)

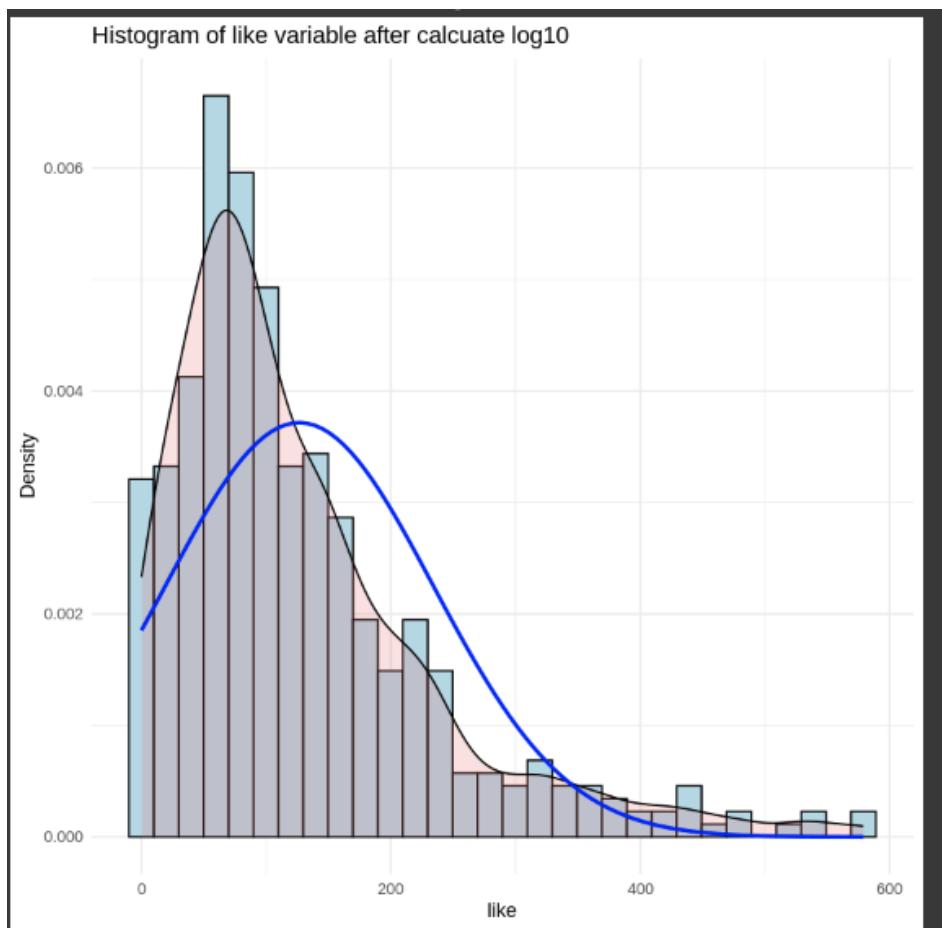
4

5 av_residual = rstandard(int_model)
6 shapiro.test(av_residual)

7

8 # Trực quan bằng QQ plot
9 qqnorm(av_residual)
10 qqline(av_residual)
11 hist(av_residual)

```



Hình 3.149: Biểu đồ sau khi loại bỏ ngoại lai

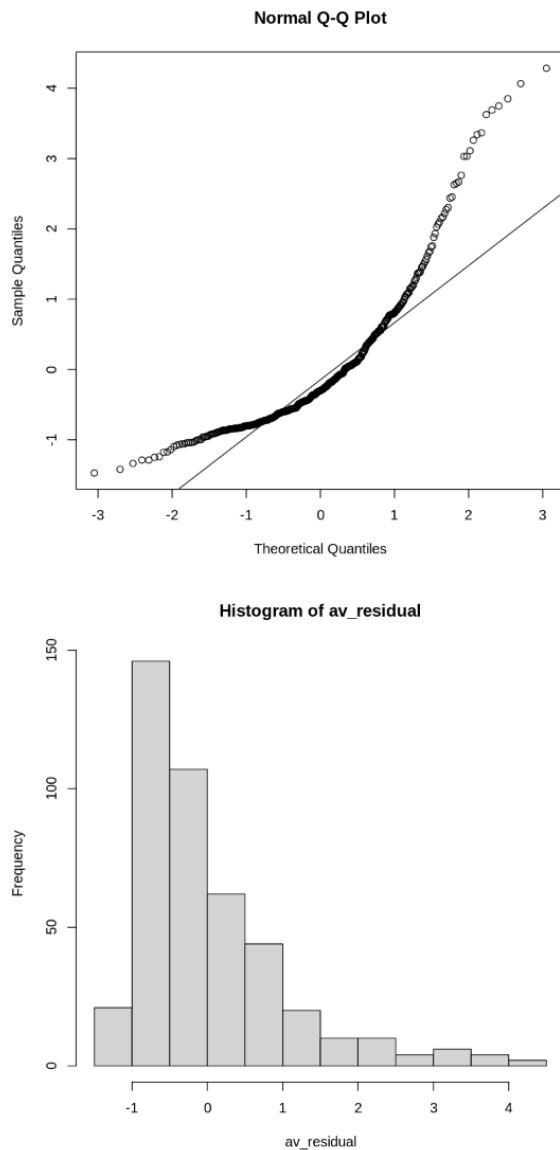
Kết quả:

```
1          Df   Sum Sq Mean Sq F value    Pr(>F)
2 Category      2   200989  100494  9.143 0.000129 ***
3 Paid          1    85534   85534  7.782 0.005512 **
4 Category:Paid 2   11079    5539  0.504 0.604487
5 Residuals     430  4726373  10992
6 ---
7 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
8 ' '
9
10        Shapiro-Wilk normality test
11
12 data: av_residual
13 W = 0.84158, p-value < 2.2e-16
```

Với giả định

- H0: Phản dư tuân theo phân phối chuẩn
- H1: Phản dư không tuân theo phân phối chuẩn

Nhận xét: Với mức ý nghĩa 5%, ta thấy Category vaf Paid có mối liên hệ mật thiết (có tương tác) tới like. ta thấy rằng phản dư không tuân theo chuẩn nhưng về tổng quan sẽ cho kết quả tốt hơn trước khi chưa xử lý dữ liệu.



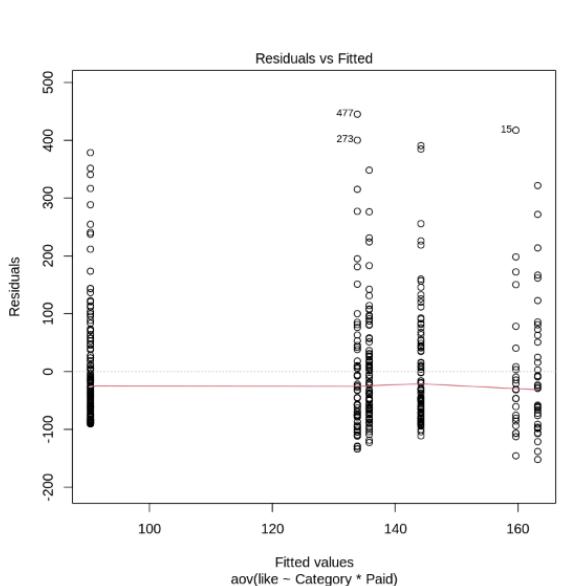
Hình 3.150: Shapiro-Wilk normality test và biểu đồ phần dư

Tiếp theo chúng ta đi kiểm định tính độc lập của phần dư:

```
1 # Kiểm định tính độc lập của phần dư
2 durbinWatsonTest(int_model)
3 plot(int_model, 1)
```

Kết quả

```
1 lag Autocorrelation D-W Statistic p-value
2     1          0.01959622      1.960124    0.678
3 Alternative hypothesis: rho != 0
```



Hình 3.151: Đồ thị Residuals

Với mức ý nghĩa 5%, ta thấy rằng mô hình không có sự tương quan (độc lập)

Tiếp tục Kiểm định các nhóm có phương sai đồng nhất hay không

```
1 # Kiểm định các nhóm có phương sai đồng nhất hay không
2 leveneTest(int_model)
```

Kết quả

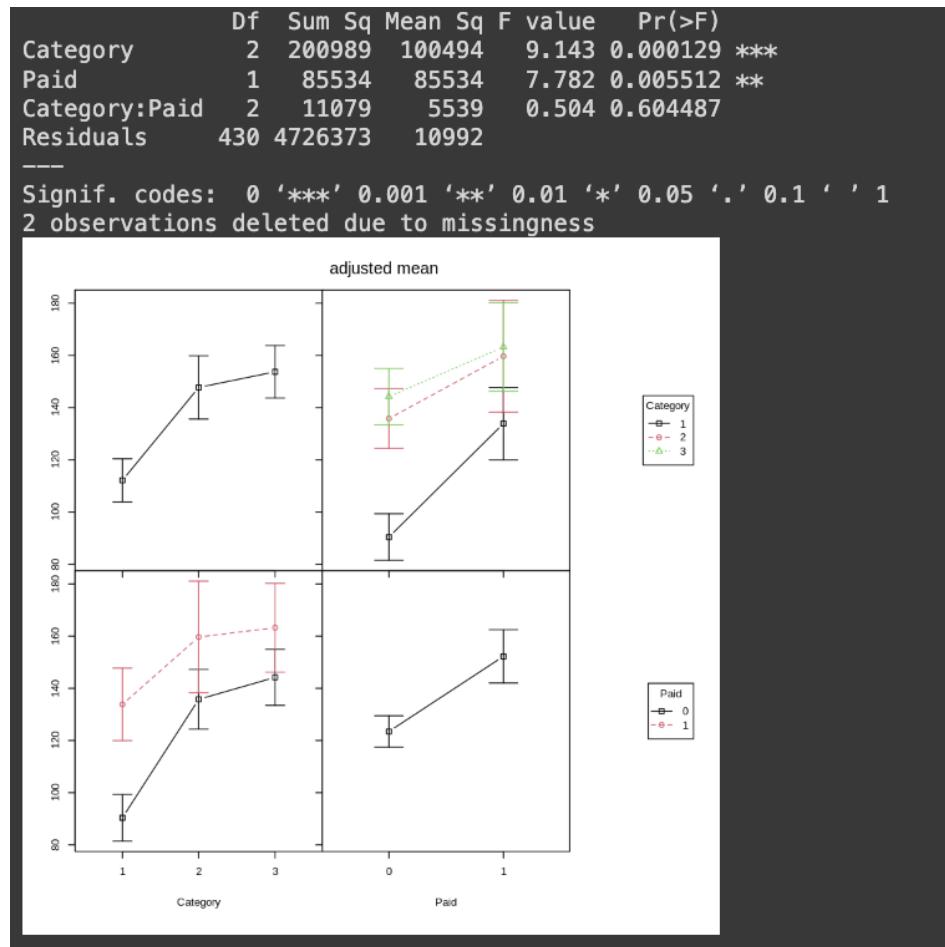
```
1 A anova: 2 x 3
2 Df      F value Pr(>F)
3 <int>    <dbl>    <dbl>
4 group     5        0.7839293   0.561649
5 430      NA       NA
```

Với mức ý nghĩa 5%, ta thấy mô hình có phương sai của các nhóm đồng nhất. Như vậy, ta đủ điều kiện để phân tích ANOVA. Bước tiếp theo, chúng ta sẽ tiến hành kiểm tra tương tác đơn và tương tác chính như phần trước.

- **Bước 1: Kiểm tra sự tương tác**

```
1 summary(int_model)
2 plot(interactionMeans(int_model))
```

Kết quả:



Hình 3.152: Tương tác giữa Category và Paid

Nhận xét:

- Với mức ý nghĩa 5% ta thấy giữa Category và Paid không có sự tương tác, nhưng bản thân chúng sẽ có sự ảnh hưởng độc lập đến like. Vì vậy ta chỉ sẽ đi phân tích ảnh hưởng chính của hai thành phần này mà không đi phân tích ảnh hưởng đơn lẻ.
- **Biểu đồ bên trái** cho thấy hiệu quả tăng dần khi thay đổi từ 1-2-3 (ảnh hưởng của 3 là rõ rệt nhất trong nhóm Category).

- **Biểu đồ bên phải** cho thấy rằng việc sử dụng quảng cáo (Paid) sẽ cho kết quả tốt hơn khi không sử dụng quảng cáo.

- **Bước 3. Phân tích ảnh hưởng chính**

- Phân tích ảnh hưởng chính của Category với hiệu quả tương tác bài post thông qua lượt like

```

1 category_model = lm(like ~ Category, data = rm_outliner_
    data)
2 summary(category_model)

```

Kết quả:

```

1
2             Call:
3 lm(formula = like ~ Category, data = rm_outliner_data)
4
5 Residuals:
6   Min     1Q Median     3Q    Max
7 -138.63 -73.09 -31.09  39.91 475.91
8
9 Coefficients:
10            Estimate Std. Error t value Pr(>|t|)
11 (Intercept) 103.092     7.551 13.653 < 2e-16 ***
12 Category2    37.541    12.610  2.977 0.003074 **
13 Category3    46.539    11.858  3.925 0.000101 ***
14 ---
15 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
16          0.1 ' ' 1
17
18 Residual standard error: 105.4 on 434 degrees of
19 freedom
20 (1 observation deleted due to missingness)
21 Multiple R-squared:  0.03975 , Adjusted R-squared:
22          0.03533
23 F-statistic: 8.984 on 2 and 434 DF, p-value: 0.0001503

```

Nhận xét: Nhận xét: Với mức ý nghĩa 0.05, ta thấy rằng các nhóm category đều ảnh hưởng đến số lượt like

```

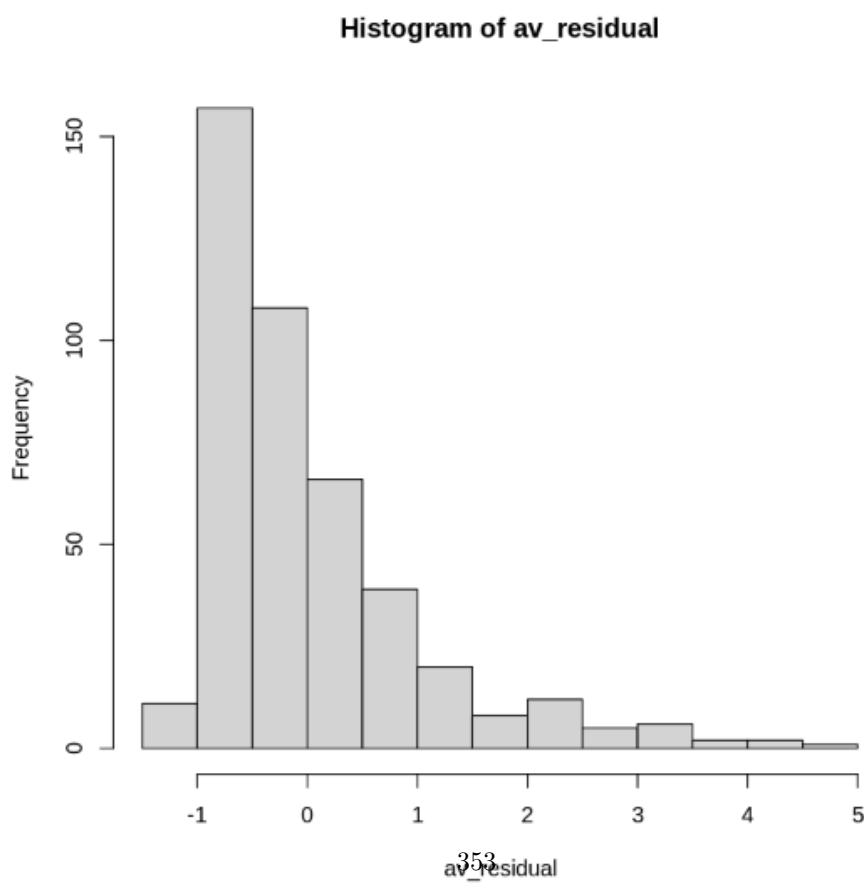
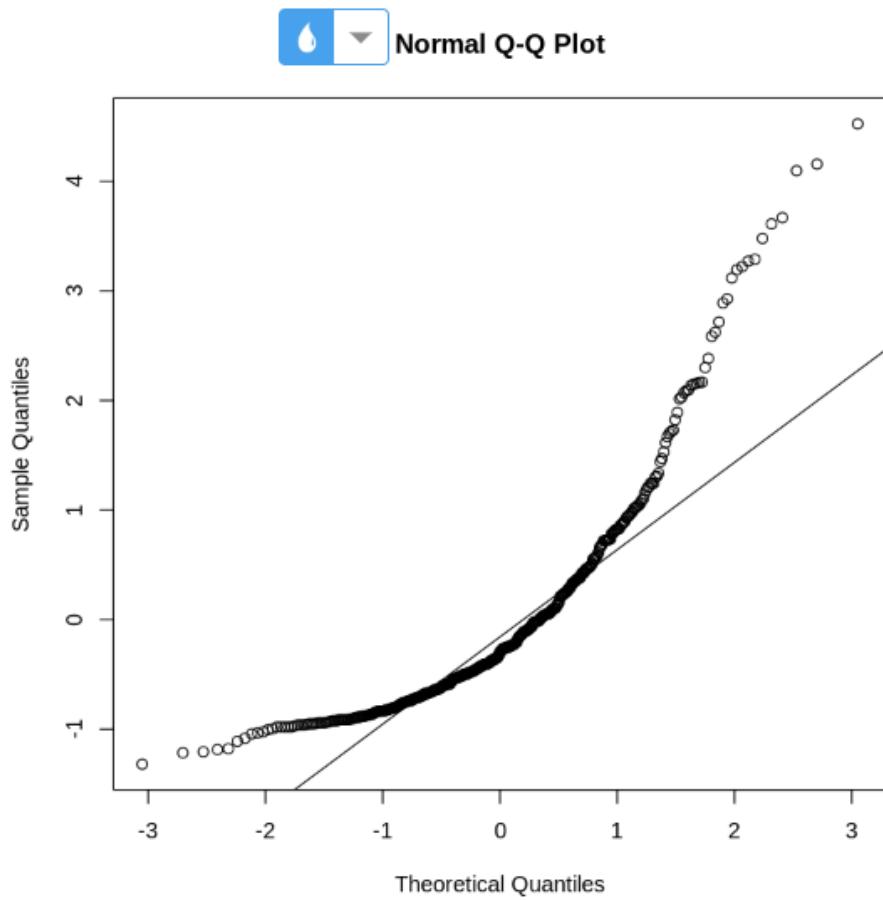
1 # Shapiro-Wilk test
2 av_residual = rstandard(category_model)

```

```
3 shapiro.test(av_residual)
4
5 # Trực quan bằng QQ plot
6 qqnorm(av_residual)
7 qqline(av_residual)
8 hist(av_residual)
```

Kết quả:

```
1 Shapiro-Wilk normality test
2
3 data: av_residual
4 W = 0.83235, p-value < 2.2e-16
```



Hình 3.153: Shapiro-test

Nhận xét: Giá trị p-value đã tăng lên rất nhiều (mặc dù < 0.05), hình dáng đồ thị gần chuẩn hơn so với trước khi chưa xử lý dữ liệu

```
1 # Kiểm định các nhóm có phương sai đồng nhất hay không
2 leveneTest(category_model)
```

Kết quả:

```
1 A anova: 2 x 3
2 Df      F value Pr(>F)
3 <int>    <dbl>    <dbl>
4 group     2        0.1707522   0.843087
5 434       NA       NA
```

Nhận xét: Với các giả định:

- H0: Các nhóm có phương sai đồng nhất
- H1: Các nhóm không có phương sai đồng nhất

Nhận xét: Với giá trị p-value = $0.843087 > 0.05$, ta không điều kiện bác bỏ H0, vậy các nhóm có phương sai đồng nhất (giá trị p-value tăng lên).

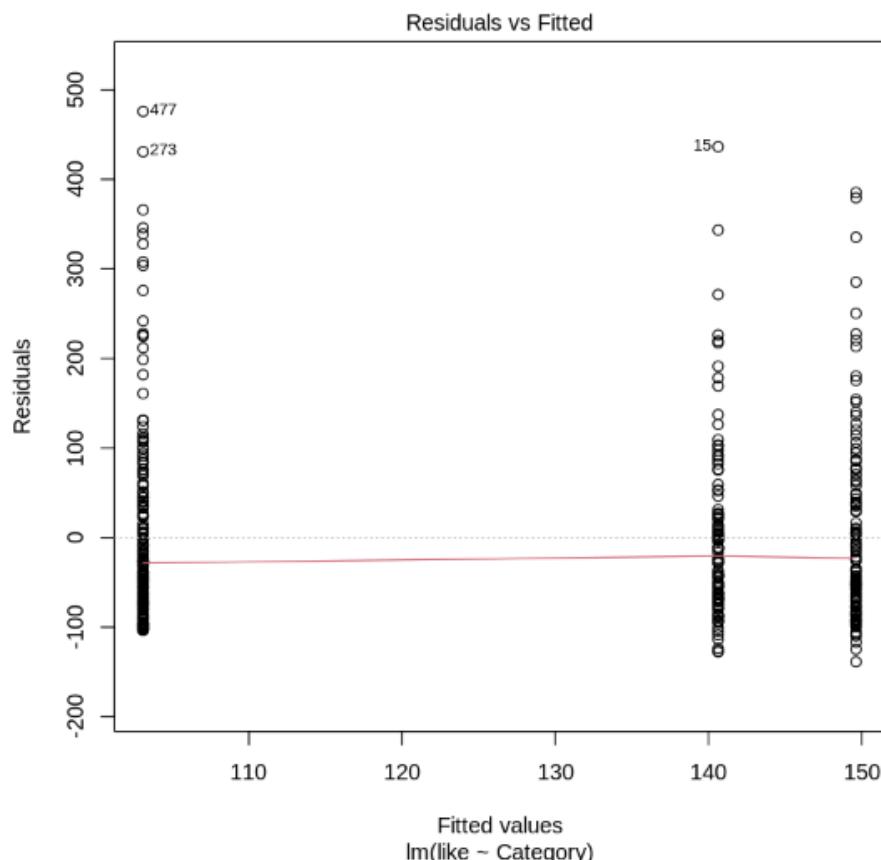
```
1 # Kiểm định tính độc lập của phần dư
2 durbinWatsonTest(category_model)
3 plot(category_model, 1)
```

Kết quả:

```

lag Autocorrelation D-W Statistic p-value
 1          0.0667957    1.865111   0.178
Alternative hypothesis: rho != 0

```



Hình 3.154: Kiểm định độc lập phần dư

Nhận xét: Với các giả định:

- H₀: Không có sự tương quan (độc lập)
- H₁: Có sự tương quan (không độc lập)

Nhận xét: Với giá trị p-value = 0.178 nên có sự tương quan dương, tuy nhiên kết quả này lớn hơn kết quả trước đó (=0).

```

1 # Kiểm định trung bình giữa các nhóm
2 with(rm_outliner_data, pairwise.t.test(like, Category, p.
   adj = "bonferroni"))
3 TukeyHSD(aov(like ~ Category, data=rm_outliner_data), conf.
   level = 0.95)

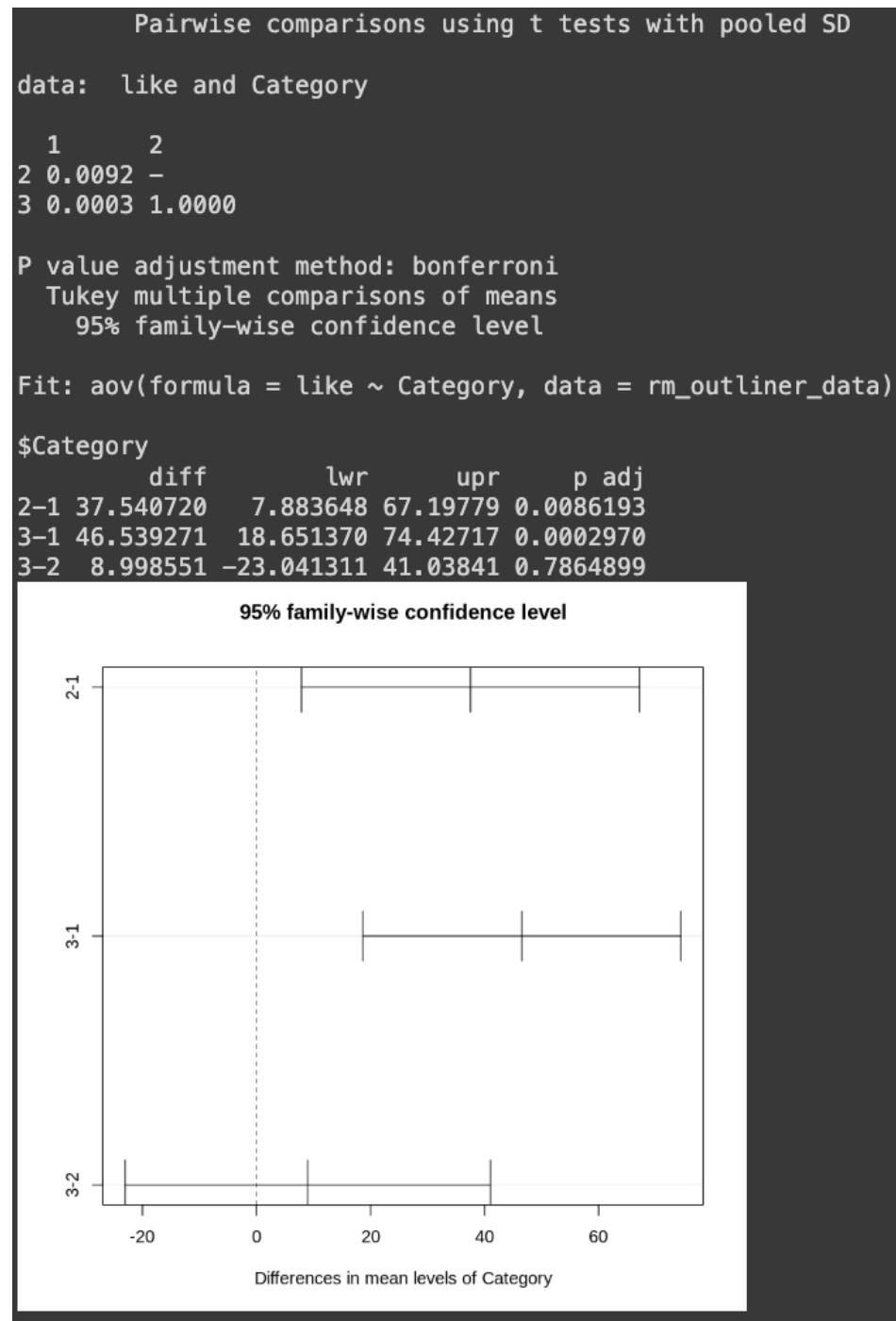
```

```

4 plot(TukeyHSD(aov(like ~ Category, data=rm_outliner_data),
                 conf.level = 0.95))

```

Kết quả:



Hình 3.155: Kiểm định trung bình

Với các giả định:

- H0: Các giá trị trung bình giữa các cặp bằng nhau
- H1: Các giá trị trung bình giữa các cặp không bằng nhau

Nhận xét:

- Cặp 2-1 và 3-1 không có mean bằng nhau (đồ thị ko cắt điểm 0, p-value > 0.05)
- Cặp 3-2 có mean bằng nhau (đồ thị cắt điểm 0, p-value < 0.05)

- Phân tích ảnh hưởng chính của Quảng cáo với hiệu quả của bài post thông qua số lượt like

```

1 paid_model = lm(like ~ Paid, data = rm_outliner_data)
2 summary(paid_model)

```

Kết quả:

```

1
2             Call:
3 lm(formula = like ~ Paid, data = rm_outliner_data)
4
5 Residuals:
6   Min     1Q Median     3Q    Max
7 -148.43 -71.43 -29.55  40.70 430.57
8
9 Coefficients:
10            Estimate Std. Error t value Pr(>|t|)
11 (Intercept) 118.546      5.996 19.770 < 2e-16 ***
12 Paid1        29.883     11.478  2.604  0.00954 **
13 ---
14 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
15
16 Residual standard error: 106.8 on 434 degrees of freedom
17   (2 observations deleted due to missingness)
18 Multiple R-squared:  0.01538 ,   Adjusted R-squared:
19                 0.01311
20 F-statistic: 6.779 on 1 and 434 DF,  p-value: 0.009542

```

Nhận xét: Với mức ý nghĩa 0.05, ta thấy rằng Paid có ý nghĩa trong việc giải thích mô hình.

```

1 # Shapiro-Wilk test
2 av_residual = rstandard(paid_model)
3 shapiro.test(av_residual)

```

```

4
5 # Trực quan bằng QQ plot
6 qqnorm(av_residual)
7 qqline(av_residual)
8 hist(av_residual)

```

Kết quả:

```

1 Shapiro-Wilk normality test
2
3 data: av_residual
4 W = 0.9859, p-value = 0.07921

```

Với các giả định:

- H0: Tuân theo phân phối chuẩn
- H1: Không tuân theo phân phối chuẩn

Nhận xét: với độ tin cậy 5% thì với giá trị p-value = 2.2e-16 chúng ta đủ cơ sở bác bỏ H0, vậy sai số có phân phối không chuẩn. (Giống kết quả trước đó), tuy nhiên về biểu đồ phân phối cho ta kết quả đẹp hơn

```

1 # Kiểm định các nhóm có phương sai đồng nhất hay không
2 leveneTest(drug_model)

```

Kết quả:

```

1 A anova: 2 x 3
2 Df      F value Pr(>F)
3 <int>    <dbl>    <dbl>
4 group     1        2.750614    0.09793961
5 434      NA       NA

```

Với các giả định:

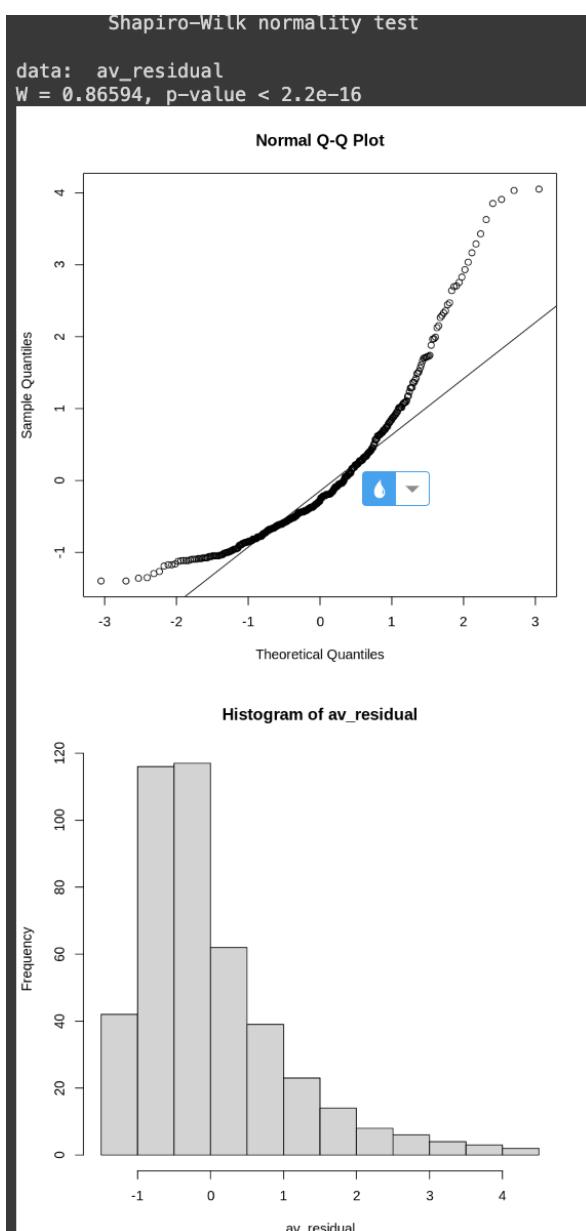
- H0: Các nhóm có phương sai đồng nhất
- H1: Các nhóm không có phương sai đồng nhất

Nhận xét: Với giá trị p-value = 0.097 > 0.05, ta không đủ điều kiện bác bỏ H0, vậy các nhóm có phương sai đồng nhất (giống cũ)

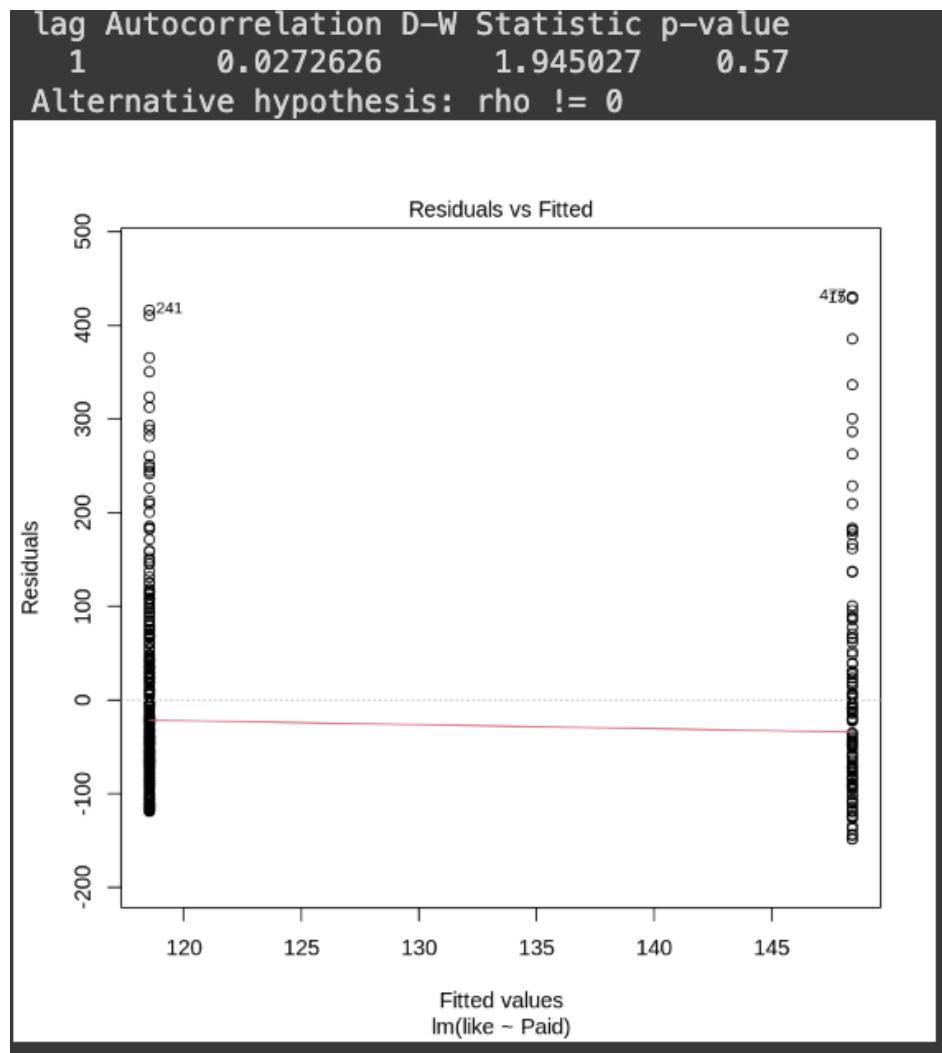
```

1 # Kiểm định tính độc lập của phần dư
2 durbinWatsonTest(paid_model)
3 plot(paid_model, 1)

```



Hình 3.156: Shapiro-Wilk normality test và độ thị phân phối



Hình 3.157: Kiểm định tính độc lập của phần dư

Với các giả định:

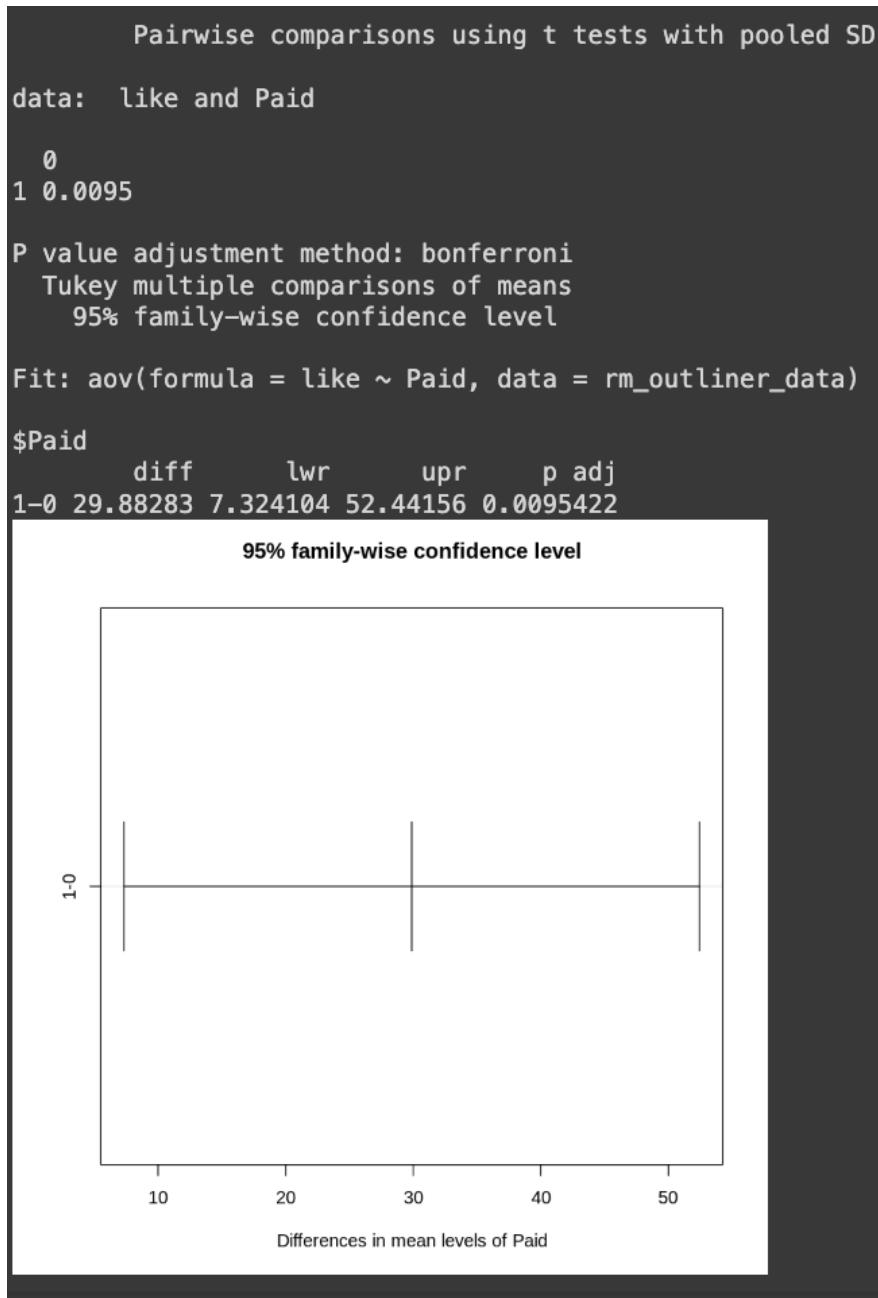
- H0: Không có sự tương quan (độc lập)
- H1: Có sự tương quan (không độc lập)

Nhận xét: Với giá trị p-value = 0.57 nên không có sự tương quan (trước đó là 0.11).

```

1 # Kiểm định độ hiệu quả trung bình
2 with(rm_outliner_data, pairwise.t.test(like, Paid, p.adj =
  "bonferroni"))
3 TukeyHSD(aov(like ~ Paid, data=rm_outliner_data), conf.level =
  0.95)
4 plot(TukeyHSD(aov(like ~ Paid, data=rm_outliner_data), conf.
  level = 0.95))

```



Hình 3.158: Kiểm định độ hiệu quả trung bình

Với các giả định:

- H₀: Các giá trị trung bình giữa các cặp bằng nhau
- H₁: Các giá trị trung bình giữa các cặp không bằng nhau

Nhận xét: Nhìn vào kết quả ta có: Với p-value = 0.095 thì ta không đủ điều kiện bác bỏ H₀, vậy 2 nhóm có mean bằng nhau.

- **Bước 4: Xây dựng và kiểm định mô hình cộng (Additive model)**

```

1 add_model = lm(like ~ ., data=rm_outliner_data)
2 add_model <- MASS::stepAIC(add_model, k = log(nrow(rm_
    outliner_data)), trace = 0)
3 summary(add_model)
4 add_model$coefficients

```

Kết quả:

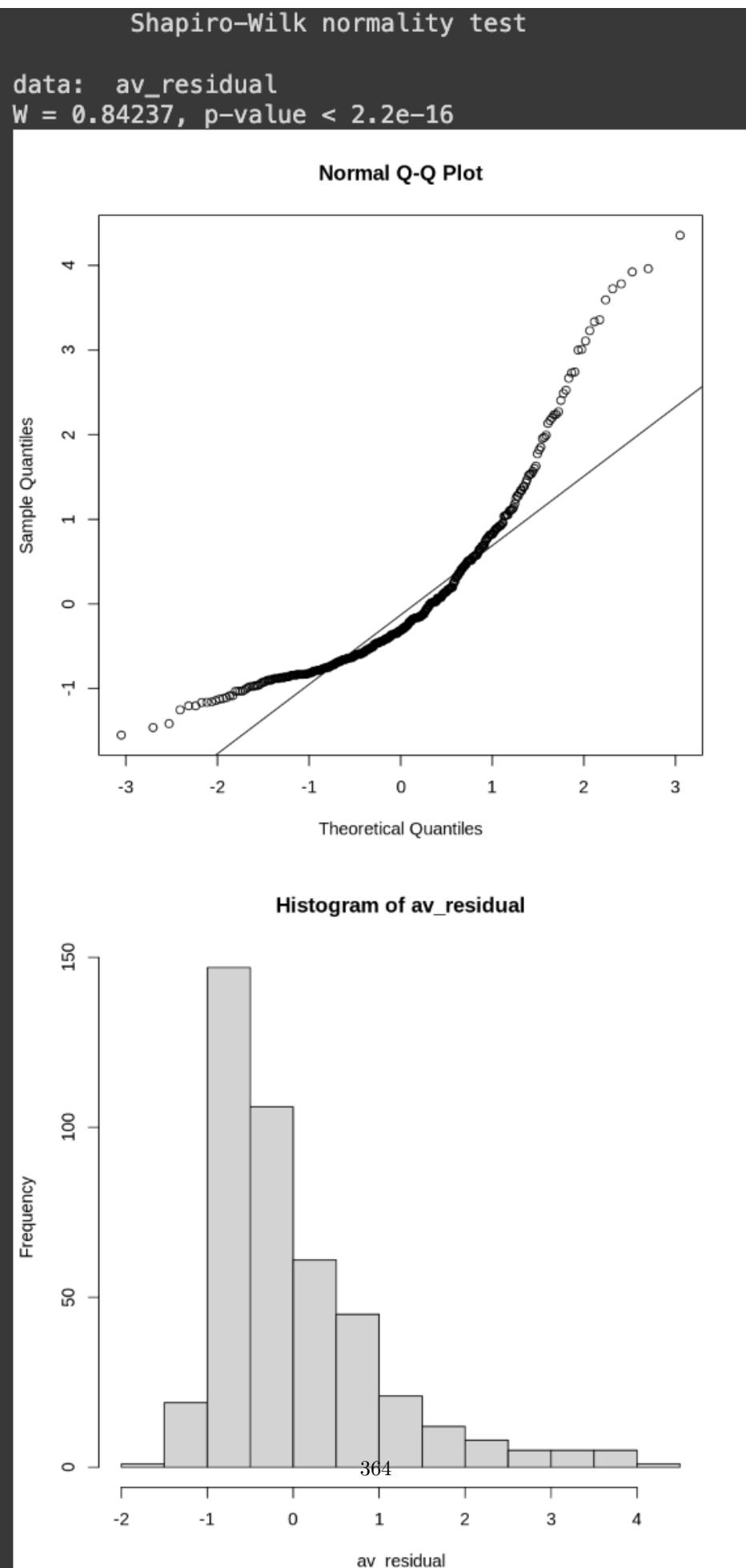
```

1
2 Call:
3 lm(formula = like ~ Category + Paid, data = rm_outliner_
    data)
4
5 Residuals:
6   Min     1Q Median     3Q    Max
7 -161.14 -71.10 -32.51  44.18 453.61
8
9 Coefficients:
10                      Estimate Std. Error t value Pr(>|t|)
11 (Intercept) 93.882      8.192 11.460 < 2e-16 ***
12 Category2    40.209     12.586  3.195 0.00150 **
13 Category3    46.747     11.777  3.969 8.44e-05 ***
14 Paid1        31.510     11.283  2.793 0.00546 **
15 ---
16 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
17 ' '
18 Residual standard error: 104.7 on 432 degrees of freedom
19   (2 observations deleted due to missingness)
20 Multiple R-squared:  0.05703 , Adjusted R-squared:
21          0.05048
22 F-statistic: 8.709 on 3 and 432 DF,  p-value: 1.275e-05
23 (Intercept) 93.8817153094315
24 Category2 40.2086725476412
25 Category3 46.746.7470289779827
26 Paid1 31.5099213098397

```

Nhận xét: Với p-value=5%, cả 2 biến đều có ý nghĩa trong việc giải thích mô hình. Ta tiến hành kiểm định Shapiro và Breusch-Pagan

```
1 # Shapiro-Wilk test
2 av_residual = rstandard(add_model)
3 shapiro.test(av_residual)
4
5 # Trực quan bằng QQ plot
6 qqnorm(av_residual)
7 qqline(av_residual)
8 hist(av_residual)
```



Nhận xét: với độ tin cậy 5% thì với giá trị p-value = 2.2e-16 chúng ta đủ cơ sở bác H₀, vậy phần dư không tuân theo chuẩn (giống trước).

```

1 # Kiểm định tính độc lập của phần dư
2 durbinWatsonTest(add_model)
3 plot(add_model, 1)
```

Kết quả:

```

1 lag Autocorrelation D-W Statistic p-value
2   1      0.02458993     1.950179    0.606
3 Alternative hypothesis: rho != 0
```

Nhận xét: Với giá trị p-value = 0.606 nên có sự tương quan dương, tuy nhiên kết quả tốt hơn trước (0.412).

```

1 # Kiểm định Breusch-Pagan
2 bptest(add_model)
3 add_model$coefficients
```

Kết quả:

```

1 studentized Breusch-Pagan test
2
3 data: add_model
4 BP = 4.3594, df = 3, p-value = 0.2252
5 (Intercept) 93.8817153094315 Category2 40.21
6 Category3 46.747 Category4 31.51 Paid1
```

Với p-value=0.2252 > 0.05 thì ta không đủ điều kiện bác bỏ H₀. Vậy phương sai của mô hình độc lập. Như vậy, mô hình cộng được xây dựng như sau:

Mô hình cộng được xây dựng như sau:

$$\text{like} = 93.38 + 40.20 \times \text{Category2} + 46.747 \times \text{Category3} + 31.51 \times \text{Paid1} \quad (3.1)$$

- **Category2 (40.21)**: Khi biến Category chuyển từ mức cơ bản (Category1) sang Category2, biến phụ thuộc tăng trung bình 40.21 đơn vị.
- **Category3 (46.75)**: Khi biến Category chuyển từ mức cơ bản (Category1) sang Category3, biến phụ thuộc tăng trung bình 46.75 đơn vị.
- **Paid1 (31.51)**: Khi biến Paid chuyển từ mức cơ bản (Paid0) sang Paid1, biến phụ thuộc tăng trung bình 31.51 đơn vị.

Kết luận: Như vậy để tăng tương tác bài viết, người dùng có thể sử dụng quảng cáo và nội dung liên quan đến chủ đề 2 và 3 (chủ đề 3 cho kết quả cao hơn).

Nhận xét chung: Như vậy về tổng thể, sau khi loại bỏ các điểm ngoại lai và cực ngoại lai, việc thống kê và phân tích ANOVA đã cho ra một mô hình có các yếu tố thỏa mãn các kiểm định về chuẩn hơn. Trong trường hợp không chuẩn, các chỉ số đã được cải thiện so với trước đây.

CHƯƠNG 4

TỔNG KẾT ĐỒ ÁN

4.1. Kết luận

Trong quá trình thực hiện phân tích ANOVA, chúng tôi đã tiến hành các bước cần thiết để đảm bảo tính chính xác và tin cậy của kết quả. Cụ thể:

- **Kiểm tra các giả định cơ bản:** Các giả định về độc lập, phân phối chuẩn của dữ liệu, và đồng nhất phương sai đã được kiểm tra kỹ lưỡng. Điều này đảm bảo rằng dữ liệu đáp ứng các điều kiện cần thiết để thực hiện ANOVA.
- **Thiết kế nghiên cứu hợp lý:** Chúng tôi đã xác định và thực hiện phân tích với số lượng nhóm và mẫu phù hợp. Sự ngẫu nhiên trong phân bổ mẫu vào các nhóm giúp giảm thiểu sai lệch và nâng cao tính chính xác của kết quả.
- **Phân tích dữ liệu:** ANOVA được thực hiện để xác định sự khác biệt có ý nghĩa giữa các nhóm. Khi phát hiện sự khác biệt, chúng tôi đã tiến hành các kiểm định hậu hoc để làm rõ cặp nhóm nào có sự khác biệt cụ thể. Việc kiểm tra tương tác giữa các yếu tố cũng được thực hiện để hiểu rõ hơn về mối quan hệ giữa các biến.
- **Xử lý ngoại lệ:** Chúng tôi đã xác định và xử lý các điểm ngoại lai, thực hiện phân tích ANOVA với và không có các ngoại lệ này để đảm bảo tính toàn vẹn của kết quả.
- **Kết quả và diễn giải:** Kết quả của phân tích ANOVA đã chỉ ra sự khác biệt có ý nghĩa giữa các nhóm nghiên cứu. Chúng tôi đã diễn giải kết quả này một cách rõ ràng, kết hợp với các số liệu về độ lớn hiệu ứng để cung cấp cái nhìn toàn diện về ảnh hưởng của các yếu tố lên biến phụ thuộc.
- **Hạn chế và đề xuất:** Dù kết quả phân tích ANOVA là hợp lý và có giá trị, chúng tôi cũng đã thảo luận về các hạn chế của nghiên cứu, chẳng hạn như kích thước mẫu và các giả định không hoàn toàn được đáp ứng. Các đề xuất cho nghiên cứu tiếp theo đã được đưa ra nhằm cải thiện tính chính xác và khả năng áp dụng của mô hình.