Classificação Antifraude em Seguros Automotivos usando Funções Discriminantes Lineares e Quadráticas

Table of contents

1 Contexto e Objetivo

Uma seguradora de automóveis quer automatizar a triagem de sinistros para decidir quais casos devem ser auditados com prioridade (investigação antifraude). Historicamente, o setor antifraude classifica sinistros em:

- **Classe A (Regular):** sinistro com perfil compatível com o padrão esperado.
- **Classe B (Suspeito):** sinistro com padrão atípico (alto risco de inconsistência/fraude).

Para cada sinistro, são coletadas variáveis operacionais e comportamentais:

- `valor_reparo (R$)`
- `dias_para_reportar`: dias entre o evento e o aviso
- `qt_sinistros_12m`: nº de sinistros nos últimos 12 meses

- `ratio_reparo_vs_fipe`: valor reparo / valor FIPE do veículo
- `dist_evento_resid`: km entre local do evento e residência
- `mudancas_contato_6m`: nº de mudanças de telefone/endereço nos últimos 6 meses

O **objetivo** desta análise é usar Análise Discriminante para classificar sinistros como Regular vs Suspeito e justificar quando usar LDA e quando usar QDA, comparando desempenho e suposições.

---

2 Pacotes utilizados

```
load <- function(pkg){
  new.pkg <- pkg[!(pkg %in% installed.packages()[, "Package"])]
  if (length(new.pkg))
    install.packages(new.pkg, dependencies = TRUE)
  sapply(pkg, require, character.only = TRUE)
}

## Pacotes utilizados nessa análise

packages = c("tidyverse", "MASS", "biotools", "caret", "pROC", "ggplot2", "Matrix", "stats", "reshape2"
load(packages)
```

```
   tidyverse         MASS     biotools        caret         pROC      ggplot2
        TRUE         TRUE         TRUE         TRUE         TRUE         TRUE
      Matrix        stats     reshape2     corrplot RColorBrewer
        TRUE         TRUE         TRUE         TRUE         TRUE
```

3 Leitura e exploração dos dados

os dados para esta análise estão disponíveis aqui.

```
df %>%
  str()
```

```
'data.frame':   400 obs. of  7 variables:
 $ valor_reparo        : num  4907 9681 5055 8929 11344 ...
 $ dias_para_reportar  : num  6.95 6.2 5.32 6.93 4.93 ...
 $ qt_sinistros_12m    : num  2 1 1 0 2 1 1 1 0 2 ...
 $ ratio_reparo_vs_fipe: num  0.169 0.502 0.538 0.147 0.315 ...
 $ dist_evento_resid   : num  8.35 24.46 29.22 19.11 17.49 ...
 $ mudancas_contato_6m : num  0 2 0 1 1 2 0 1 0 0 ...
 $ classe              : Factor w/ 2 levels "Regular","Suspeito": 1 1 1 1 1 1 1 1 1 1 ...
```

```
df %>%
  summary()
```

```
 valor_reparo   dias_para_reportar qt_sinistros_12m ratio_reparo_vs_fipe
Min.   : 2731   Min.   : 1.229     Min.   :0.0000   Min.   :0.02225
1st Qu.: 6336   1st Qu.: 5.304     1st Qu.:0.0000   1st Qu.:0.31351
Median : 7995   Median : 7.223     Median :1.0000   Median :0.48738
Mean   : 8449   Mean   : 7.890     Mean   :0.8025   Mean   :0.53469
3rd Qu.: 9553   3rd Qu.: 9.758     3rd Qu.:1.0000   3rd Qu.:0.70367
Max.   :26571   Max.   :19.550     Max.   :5.0000   Max.   :1.46809
dist_evento_resid mudancas_contato_6m      classe
Min.   :  2.682   Min.   :0.0000      Regular :320
1st Qu.: 15.873   1st Qu.:0.0000      Suspeito: 80
Median : 25.454   Median :0.0000
Mean   : 29.362   Mean   :0.5125
3rd Qu.: 37.239   3rd Qu.:1.0000
Max.   :134.776   Max.   :4.0000
```

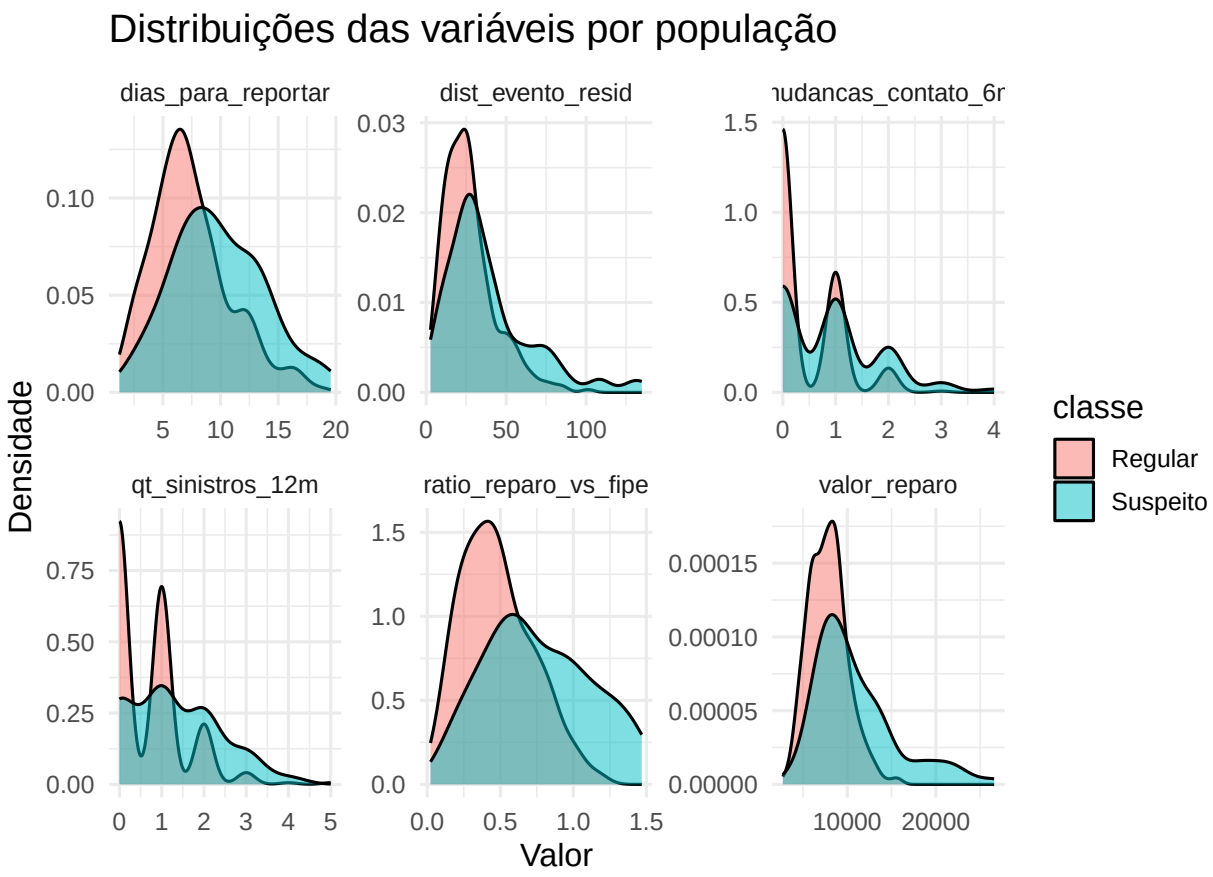
O conjunto de dados analisado é composto por 400 observações e 7 variáveis, sendo a variável resposta `classe` formada por dois grupos: `Regular` (n = 320) e `Suspeito` (n = 80), indicando uma amostra desbalanceada com predominância de sinistros regulares. As variáveis explicativas representam características operacionais dos sinistros e apresentam escalas distintas, com valores centrados próximos de zero em alguns casos, sugerindo possível padronização prévia.

De forma geral, observa-se que a população `Suspeito` tende a apresentar valores mais elevados em indicadores associados a risco, como `dias_para_reportar`, `qt_sinistros_12m`, `ratio_reparo_vs_fipe`, `dist_evento_resid` e `mudancas_contato_6m`. As medidas-resumo indicam maior dispersão em algumas variáveis, especialmente em `dist_evento_resid` e `ratio_reparo_vs_fipe`, o que sugere heterogeneidade estrutural entre as classes. Além disso, a presença de amplitudes elevadas e possíveis valores extremos reforça a necessidade de técnicas multivariadas que considerem diferenças de variância e covariância entre os grupos.

Esses padrões descritivos iniciais indicam potencial separação entre as populações, mas também evidenciam diferenças na variabilidade interna dos grupos, aspecto relevante para a escolha entre funções discriminantes lineares (LDA) e quadráticas (QDA) nas etapas posteriores da análise.

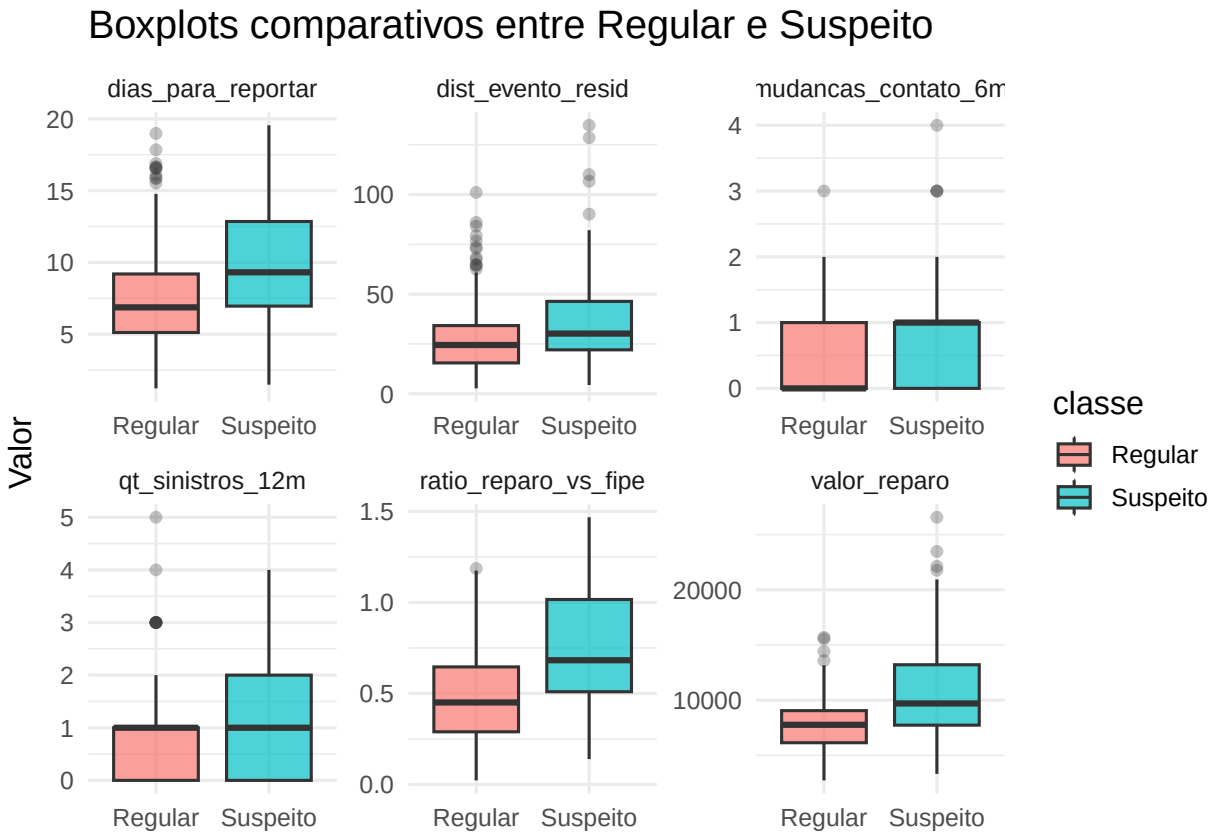
```r
df_long <- df %>%
  pivot_longer(-classe, names_to = "variavel", values_to = "valor")

ggplot(df_long, aes(x = valor, fill = classe)) +
  geom_density(alpha = 0.5) +
  facet_wrap(~ variavel, scales = "free", ncol = 3) +
  theme_minimal(base_size = 13) +
  labs(title = "Distribuições das variáveis por população",
       x = "Valor", y = "Densidade")
```



As distribuições das variáveis indicam diferenças claras entre as populações `Regular` e `Suspeito`, especialmente em termos de localização e dispersão. De modo geral, observa-se que os sinistros classificados como `Suspeitos` apresentam valores mais elevados em variáveis associadas ao risco operacional, como `dias_para_reportar`, `dist_evento_resid`, `qt_sinistros_12m` e `valor_reparo`, evidenciando deslocamento das densidades para a direita em comparação ao grupo `Regular`. Além disso, o grupo `Suspeito` apresenta maior variabilidade em diversas variáveis, sugerindo heterogeneidade estrutural entre as classes. Apesar da separação parcial observada, ainda há regiões de sobreposição entre as densidades, indicando que a distinção entre os grupos não é perfeitamente linear. Esse padrão visual sugere que abordagens discriminantes quadráticas podem capturar melhor diferenças na estrutura de variância e covariância entre as populações, embora modelos lineares ainda possam fornecer uma boa aproximação inicial.

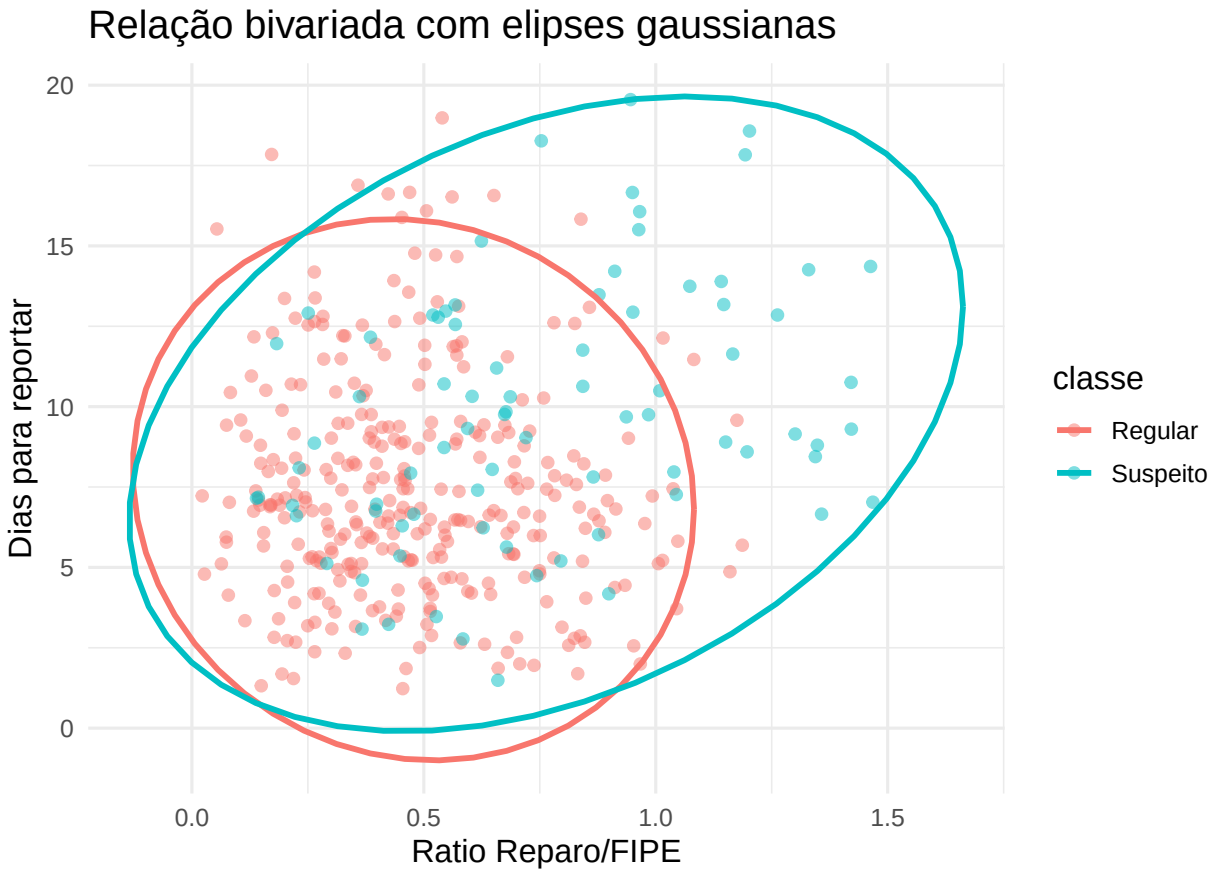
```
ggplot(df_long, aes(x = classe, y = valor, fill = classe)) +
  geom_boxplot(alpha = 0.7, outlier.alpha = 0.3) +
  facet_wrap(~ variavel, scales = "free", ncol = 3) +
  theme_minimal(base_size = 13) +
  labs(title = "Boxplots comparativos entre Regular e Suspeito",
       x = "", y = "Valor")
```



Os boxplots evidenciam diferenças sistemáticas entre as classes `Regular` e `Suspeito`, tanto em termos de posição quanto de dispersão das variáveis. Observa-se que a população `Suspeito` apresenta medianas consistentemente mais elevadas em `dias_para_reportar`, `dist_evento_resid`, `mudancas_contato_6m`, `qt_sinistros_12m` e `valor_reparo`, indicando maior frequência de características associadas a padrões atípicos de sinistro. Além disso, a amplitude interquartílica e a presença de valores extremos são mais pronunciadas no grupo `Suspeito`, sugerindo maior variabilidade interna. A variável `ratio_reparo_vs_fipe` apresenta maior sobreposição entre as classes, indicando menor poder discriminante isolado.

De modo geral, embora exista separação clara entre os grupos em várias variáveis, a presença de sobreposição e diferenças na dispersão reforça a hipótese de heterogeneidade nas matrizes de covariância entre as populações, aspecto relevante para a escolha entre funções discriminantes lineares e quadráticas nas etapas subsequentes da análise.

```
ggplot(df, aes(x = ratio_reparo_vs_fipe,
               y = dias_para_reportar,
               color = classe)) +
  geom_point(alpha = 0.5) +
  stat_ellipse(type = "norm", linewidth = 1.1) +
  theme_minimal(base_size = 13) +
  labs(title = "Relação bivariada com elipses gaussianas",
       x = "Ratio Reparo/FIPE",
       y = "Dias para reportar")
```

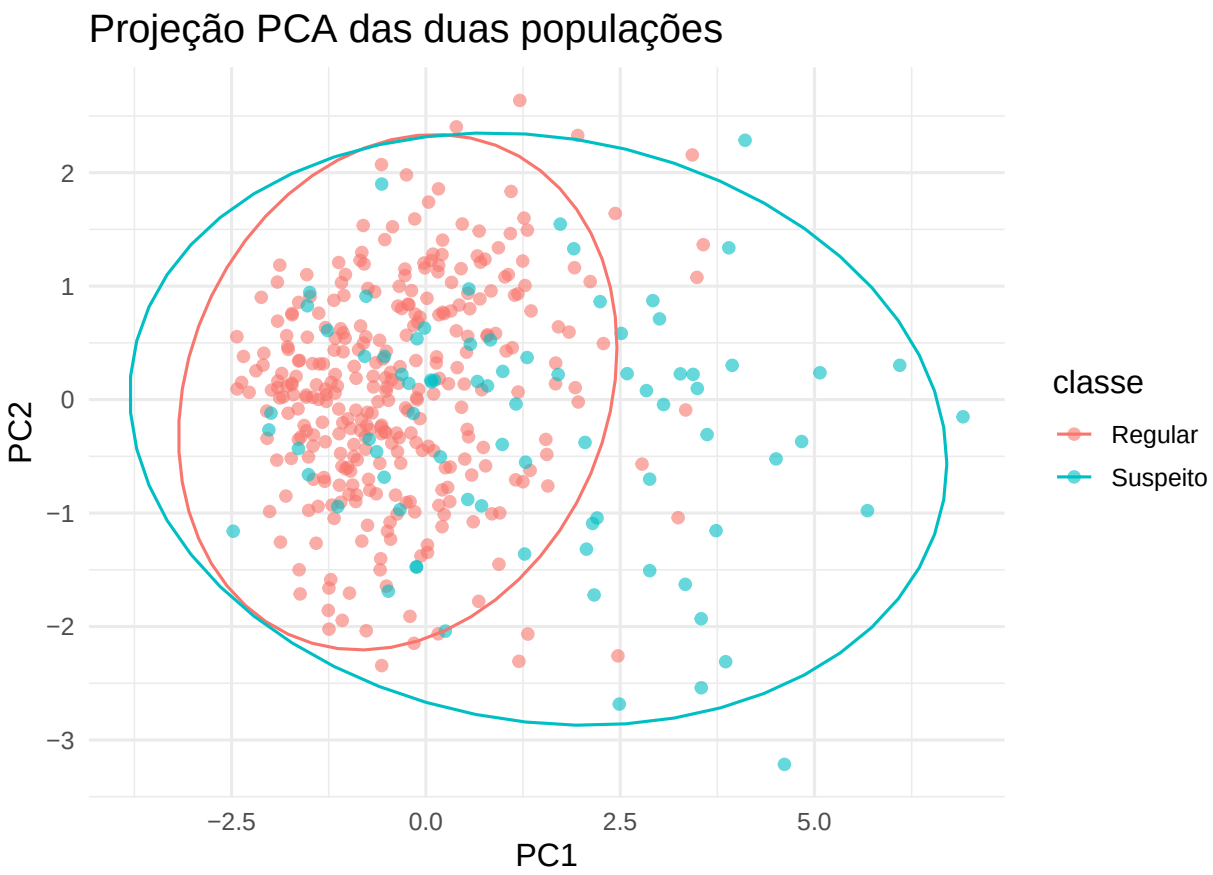


A relação bivariada entre `ratio_reparo_vs_fipe` e `dias_para_reportar` revela padrões distintos entre as populações `Regular` e `Suspeito`. Observa-se que o grupo `Suspeito` apresenta maior dispersão e uma elipse de confiança significativamente mais ampla, indicando variabilidade superior e possível estrutura de covariância diferente em relação ao grupo `Regular`. Além disso, há um deslocamento vertical das observações `Suspeitas` para valores mais elevados de `dias_para_reportar`, sugerindo contribuição discriminante dessa variável. A diferença no formato e na orientação das elipses gaussianas sugere heterogeneidade nas matrizes de covariância entre as classes, indicando que fronteiras de decisão não lineares podem ser mais adequadas, o que reforça a potencial vantagem do uso de funções discriminantes quadráticas em relação ao modelo linear de Fisher.

```
X_scaled <- scale(df[, -7])
pca <- prcomp(X_scaled)

pca_df <- data.frame(pca$x[,1:2], classe = df$classe)

ggplot(pca_df, aes(PC1, PC2, color = classe)) +
  geom_point(alpha = 0.6) +
  stat_ellipse(type = "norm") +
  theme_minimal(base_size = 13) +
  labs(title = "Projeção PCA das duas populações",
       x = "PC1", y = "PC2")
```



A projeção das observações nas duas primeiras componentes principais permite analisar a estrutura geométrica dos dados a partir da matriz de covariância conjunta. Como a PCA identifica direções de máxima variabilidade global, o alongamento observado na população `Suspeito` ao longo da primeira componente principal (PC1) sugere maior dispersão multivariada nesse grupo. Esse comportamento está diretamente relacionado às diferenças nas matrizes de covariância entre as classes, refletidas no tamanho e na orientação das elipses gaussianas.

Do ponto de vista geométrico, a regra discriminante linear de Fisher (LDA) assume que as populações compartilham a mesma matriz de covariância, implicando que a separação ótima ocorre por meio de uma única direção linear no espaço das variáveis. Entretanto, a PCA revela que as nuvens de pontos apresentam formatos distintos, indicando heterogeneidade estrutural que viola parcialmente essa suposição. Quando as elipses possuem orientações e escalas diferentes, a fronteira de decisão ótima tende a ser curva, o que justifica o uso de funções discriminantes quadráticas (QDA).

Assim, a análise exploratória via PCA fornece evidências visuais de que a variabilidade do grupo `Suspeito` não apenas é maior, mas também ocorre em direções específicas do espaço multivariado, sugerindo que modelos baseados em covariâncias distintas podem capturar melhor a estrutura dos dados. Consequentemente, espera-se que o QDA apresente maior flexibilidade na delimitação das regiões de decisão, enquanto o LDA atua como uma aproximação linear potencialmente adequada quando a sobreposição entre as classes é moderada.

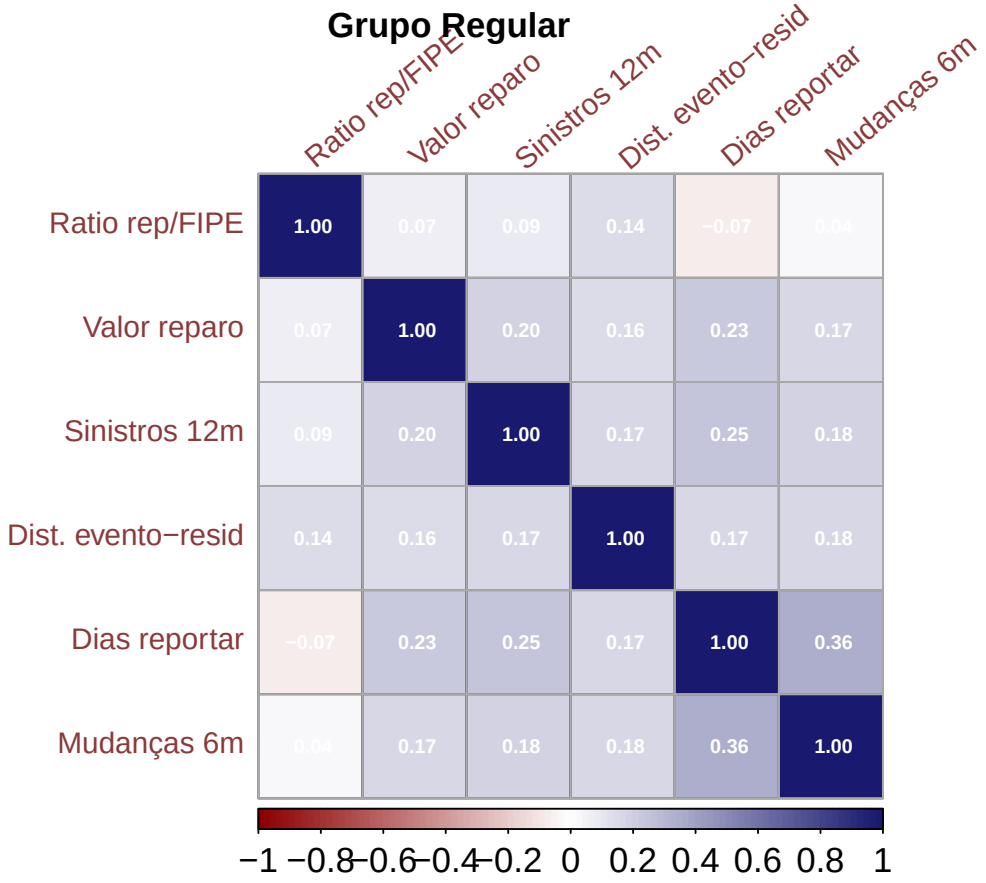
```
corr_reg <- cor(df[df$classe=="Regular",-7])
corr_sus <- cor(df[df$classe=="Suspeito",-7])

labels_vars <- c(
  valor_reparo = "Valor reparo",
```

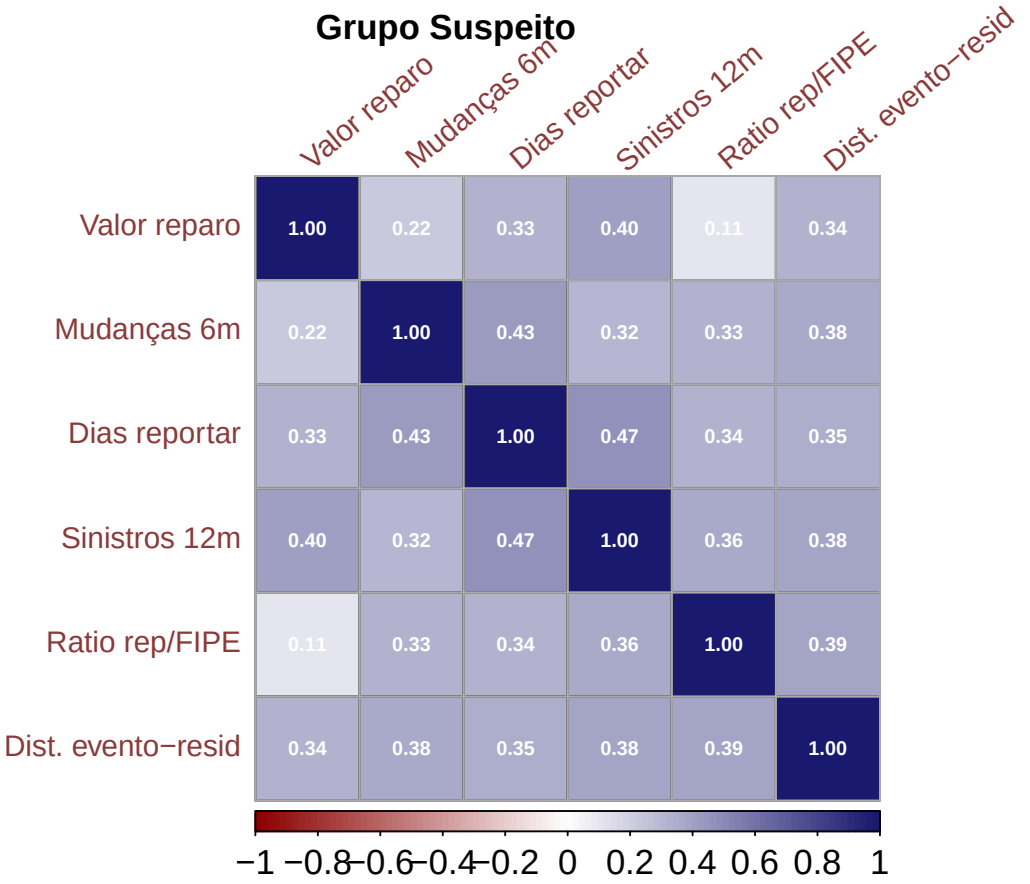
```
  dias_para_reportar = "Dias reportar",
  qt_sinistros_12m = "Sinistros 12m",
  ratio_reparo_vs_fipe = "Ratio rep/FIPE",
  dist_evento_resid = "Dist. evento-resid",
  mudancas_contato_6m = "Mudanças 6m"
)

colnames(corr_reg) <- rownames(corr_reg) <- labels_vars[colnames(corr_reg)]
colnames(corr_sus) <- rownames(corr_sus) <- labels_vars[colnames(corr_sus)]

corrplot(corr_reg,method = "color", outline = T, addgrid.col = "darkgray", order = "hclust", cl.pos = "
mtext("Grupo Regular", side = 3, line = 3, cex = 1.2, font = 2)
```




```
corrplot(corr_sus, method = "color", outline = T, addgrid.col = "darkgray", order = "hclust", cl.pos =
mtext("Grupo Suspeito", side = 3, line = 3, cex = 1.2, font = 2)
```



A análise das matrizes de correlação evidencia diferenças relevantes na estrutura de dependência entre as variáveis nos grupos `Regular` e `Suspeito`.

No grupo `Regular`, observa-se uma estrutura de correlações moderada e mais homogênea, sem padrões extremos de associação. Destacam-se correlações positivas relativamente fortes entre `dias para reportar`, `sinistros nos últimos 12 meses` e `mudanças de contato`, sugerindo que clientes com maior histórico recente de eventos tendem a apresentar comportamento mais ativo ou instável. As variáveis financeiras (`valor do reparo` e `ratio reparo/FIPE`) também apresentam associação positiva relevante, indicando consistência entre o custo absoluto do reparo e sua proporção em relação ao valor do veículo. De forma geral, a estrutura de dependência no grupo `Regular` é mais difusa, sem concentração excessiva em um único bloco de variáveis.

Em contraste, o grupo `Suspeito` apresenta correlações substancialmente mais elevadas, especialmente entre variáveis comportamentais e de histórico de sinistros. Observa-se forte associação entre `mudanças de contato`, `sinistros 12m` e `distância evento-residência`, além de correlação elevada com `dias para reportar`. Esse padrão sugere maior interdependência entre características operacionais e comportamentais, indicando possível presença de perfis mais heterogêneos e com maior variabilidade estrutural. As correlações mais intensas neste grupo apontam para uma estrutura de covariância distinta daquela observada no grupo `Regular`.

De modo geral, os resultados indicam que os dois grupos apresentam padrões de correlação diferentes, o que reforça a possibilidade de matrizes de covariância não equivalentes entre as populações. Esse achado é particularmente relevante no contexto da análise discriminante, pois sugere que modelos baseados em funções quadráticas (QDA), que permitem covariâncias distintas entre classes, podem capturar melhor a estrutura dos dados em comparação às funções lineares de Fisher (LDA), que assumem homogeneidade das matrizes de covariância.

4 Diagnóstico de suposições: covariâncias iguais?

4.1 Teste de Box (Box's M)

```
#H_0: Sigma_R = Sigma_S = Sigma

box_res <- biotools::boxM(df[, -7], df$classe)
box_res
```

```
    Box's M-test for Homogeneity of Covariance Matrices

data:  df[, -7]
Chi-Sq (approx.) = 175.44, df = 21, p-value < 2.2e-16
```

O teste de Box (Box's M) foi aplicado para avaliar a hipótese de igualdade das matrizes de covariância entre os grupos Regular e Suspeito. O resultado obtido ($\chi \approx 846,77; gl = 21; p < 2,2 \times 10^{-16}$) indica rejeição altamente significativa da hipótese nula, sugerindo que as estruturas de variância e covariância diferem entre as populações analisadas.

Sob a perspectiva da análise discriminante, esse resultado possui implicações diretas, pois a Análise Discriminante Linear (LDA) assume homogeneidade das matrizes de covariância entre classes, enquanto a Análise Discriminante Quadrática (QDA) permite estruturas específicas por grupo. Assim, a evidência empírica obtida favorece o uso de modelos mais flexíveis, capazes de acomodar heterogeneidade estrutural.

Entretanto, a interpretação do teste de Box deve ser realizada com cautela. Esse teste é conhecido por ser altamente sensível ao tamanho amostral e a desvios da normalidade multivariada. Em amostras grandes, pequenas diferenças nas matrizes de covariância podem resultar em rejeições estatisticamente significativas, mesmo quando as discrepâncias práticas são moderadas. Além disso, violações da suposição de normalidade podem inflar a estatística do teste, aumentando a probabilidade de rejeição da hipótese nula.

Dessa forma, a decisão metodológica não deve se basear exclusivamente no resultado do Box's M, mas sim em uma análise conjunta que inclua inspeção gráfica das estruturas de correlação, avaliação do desempenho preditivo dos modelos e análise da estabilidade das fronteiras de decisão. No presente estudo, a combinação entre diferenças visuais nas matrizes de correlação e o resultado altamente significativo do teste de Box reforça a evidência de que os grupos apresentam padrões de dependência distintos, sustentando a utilização de abordagens discriminantes quadráticas.

4.2 Comparação direta das covariâncias

```
cov_reg <- cov(df[df$classe == "Regular", -7])
cov_sus <- cov(df[df$classe == "Suspeito", -7])

cov_reg
```

```
                  valor_reparo dias_para_reportar qt_sinistros_12m
valor_reparo         4.732656e+06      1736.49394848     349.33459958
dias_para_reportar   1.736494e+03        11.72548018       0.70813084
qt_sinistros_12m     3.493346e+02         0.70813084       0.67319749
ratio_reparo_vs_fipe 3.480007e+01        -0.06116438       0.01911089
dist_evento_resid    5.652570e+03         9.72350496       2.32919399
```

```
mudancas_contato_6m  2.274612e+02         0.75419196       0.09149687
                     ratio_reparo_vs_fipe dist_evento_resid mudancas_contato_6m
valor_reparo                34.800074938      5652.5702264        2.274612e+02
dias_para_reportar          -0.061164381         9.7235050        7.541920e-01
qt_sinistros_12m             0.019110886         2.3291940        9.149687e-02
ratio_reparo_vs_fipe         0.060525485         0.5696093        5.567377e-03
dist_evento_resid            0.569609344       265.0373673        1.771633e+00
mudancas_contato_6m          0.005567377         1.7716332        3.825921e-01
```

```
cov_sus
```

```
                     valor_reparo dias_para_reportar qt_sinistros_12m
valor_reparo         2.157630e+07       6123.0564606     2012.9053037
dias_para_reportar   6.123056e+03         15.6636780        2.0153236
qt_sinistros_12m     2.012905e+03          2.0153236        1.1580696
ratio_reparo_vs_fipe 1.817955e+02          0.4813671        0.1412292
dist_evento_resid    4.312760e+04         38.5516270       11.2819783
mudancas_contato_6m  9.591896e+02          1.5553973        0.3117089
                     ratio_reparo_vs_fipe dist_evento_resid mudancas_contato_6m
valor_reparo                181.7954627      43127.600941         959.1895894
dias_para_reportar            0.4813671         38.551627           1.5553973
qt_sinistros_12m              0.1412292         11.281978           0.3117089
ratio_reparo_vs_fipe          0.1298240          3.879459           0.1086049
dist_evento_resid             3.8794586        761.116194           9.5362481
mudancas_contato_6m           0.1086049          9.536248           0.8449367
```

A comparação entre as matrizes de covariância dos grupos `Regular` e `Suspeito` evidencia diferenças substanciais na dispersão e na estrutura de dependência entre as variáveis. De modo geral, o grupo `Suspeito` apresenta variâncias e covariâncias significativamente maiores, indicando maior heterogeneidade e amplitude dos dados em relação ao grupo `Regular`.

No grupo `Regular`, observa-se uma estrutura de variabilidade mais moderada e relativamente estável. As covariâncias entre as variáveis comportamentais (`dias para reportar`, `sinistros nos últimos 12 meses` e `mudanças de contato`) são positivas, porém de magnitude controlada, sugerindo padrões consistentes de relacionamento sem extremos de variabilidade. Além disso, embora existam associações relevantes envolvendo variáveis financeiras e geográficas, a matriz apresenta valores mais equilibrados, refletindo um comportamento menos disperso.

Em contraste, o grupo `Suspeito` apresenta aumentos expressivos nas variâncias individuais, especialmente em `valor do reparo` e `distância evento-residência`, e covariâncias mais elevadas entre múltiplas variáveis. Destacam-se valores elevados envolvendo `ratio reparo/FIPE`, `distância evento-residência` e `sinistros 12m`, indicando que, nesse grupo, as variáveis estão mais fortemente inter-relacionadas e apresentam maior variabilidade conjunta. Esse padrão sugere a presença de maior heterogeneidade estrutural e possíveis subperfís comportamentais dentro da população classificada como suspeita.

Essas diferenças reforçam a evidência já apontada pelo teste de Box (Box's M), indicando que as matrizes de covariância não podem ser consideradas homogêneas entre os grupos. Do ponto de vista metodológico, tal resultado sugere que abordagens discriminantes que assumem covariâncias iguais, como a Análise Discriminante Linear (LDA), podem não capturar adequadamente a estrutura dos dados. Em contrapartida, modelos que permitem matrizes de covariância específicas por classe, como a Análise Discriminante Quadrática (QDA), tendem a ser mais adequados para representar a dinâmica observada entre as variáveis.

---

5 Ajuste dos modelos: LDA (Fisher) e QDA

5.1 Método da substituição

```r
X_scaled <- scale(df[, -7])
df_scaled <- data.frame(classe = df$classe, X_scaled)

fit_lda <- MASS::lda(classe ~ ., data = df_scaled)
fit_qda <- MASS::qda(classe ~ ., data = df_scaled)

fit_lda
```

```
Call:
lda(classe ~ ., data = df_scaled)

Prior probabilities of groups:
 Regular Suspeito
     0.8      0.2

Group means:
         valor_reparo dias_para_reportar qt_sinistros_12m ratio_reparo_vs_fipe
Regular    -0.1990037         -0.1289575       -0.1269541           -0.1936770
Suspeito    0.7960148          0.5158298        0.5078163            0.7747081
         dist_evento_resid mudancas_contato_6m
Regular         -0.1227236          -0.1273873
Suspeito         0.4908943           0.5095490

Coefficients of linear discriminants:
                            LD1
valor_reparo         0.60924975
dias_para_reportar   0.18151359
qt_sinistros_12m     0.07438289
ratio_reparo_vs_fipe 0.64072925
dist_evento_resid    0.04676077
mudancas_contato_6m  0.15616528
```

5.1.1 Análise Discriminante Linear (LDA)

Após a padronização das variáveis preditoras, a função discriminante linear estimada permite avaliar diretamente a contribuição relativa de cada variável para a separação entre os grupos `Regular` e `Suspeito`. Como as variáveis foram transformadas para média zero e variância unitária, a magnitude dos coeficientes passa a refletir a importância discriminante de cada atributo em uma escala comum.

Observa-se que as maiores contribuições para o eixo discriminante LD1 estão associadas às variáveis `ratio reparo/FIPE`, `valor do reparo` e `distância evento-residência`, que apresentam coeficientes de maior magnitude absoluta. Em particular, o coeficiente negativo elevado de `ratio reparo/FIPE` indica que essa variável desempenha papel central na orientação do eixo discriminante, funcionando como um fator de contraste em relação às demais variáveis comportamentais. Já `sinistros 12m` e `mudanças de contato` apresentam coeficientes positivos moderados, sugerindo que níveis mais elevados dessas características contribuem para deslocar as observações em direção ao grupo `Suspeito` no espaço discriminante.

As médias padronizadas por grupo mostram que o grupo `Suspeito` possui escores positivos na maioria das variáveis, enquanto o grupo `Regular` apresenta valores negativos, indicando que o eixo LD1 representa

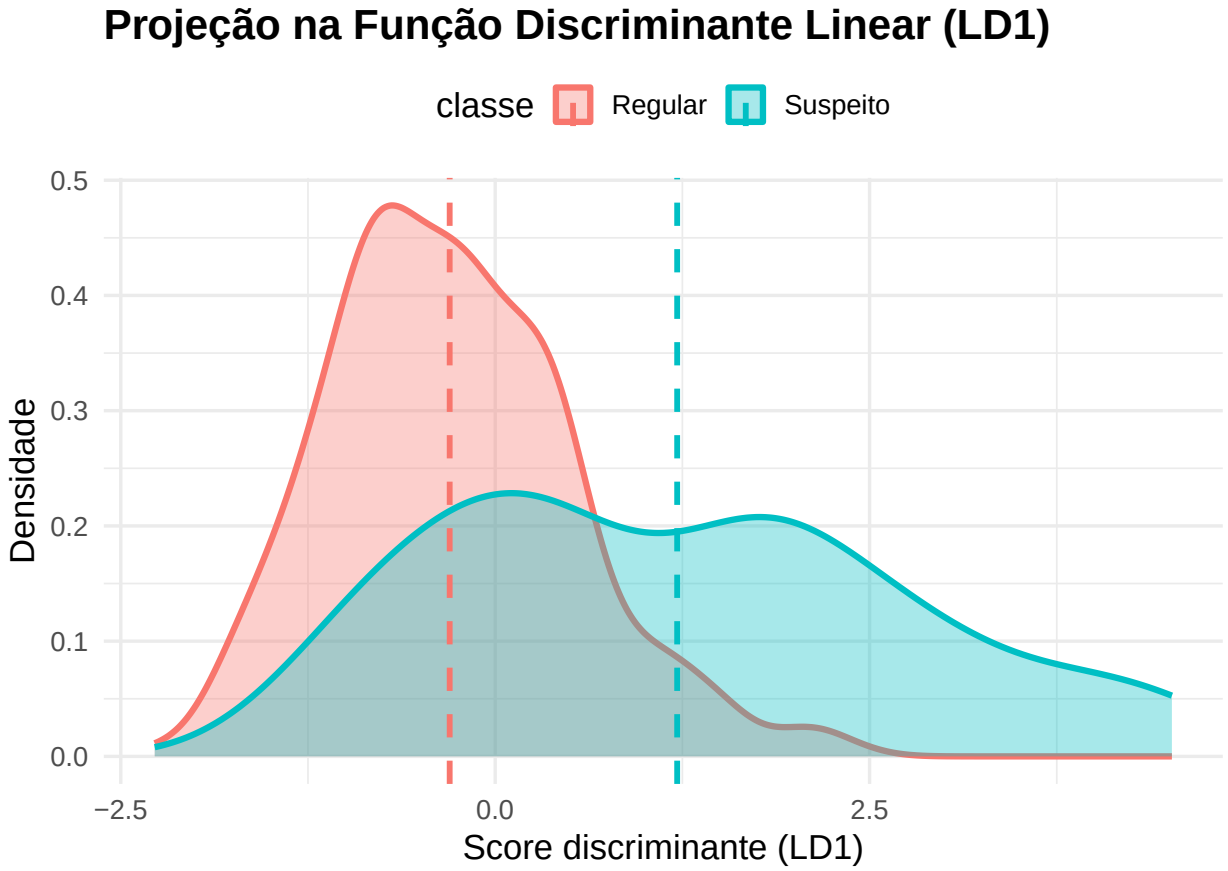
essencialmente um gradiente de intensidade operacional e comportamental. Dessa forma, indivíduos com maiores valores padronizados de histórico de sinistros, instabilidade cadastral e distância geográfica tendem a posicionar-se mais próximos da região associada ao grupo `Suspeito`.

A padronização também evidencia que a separação entre os grupos é fortemente influenciada por um conjunto integrado de variáveis, e não por um único atributo isolado, reforçando a interpretação geométrica do LDA como uma combinação linear que maximiza a separação multivariada.

```
# Obter projeções LD1
lda_proj <- predict(fit_lda)$x

dados_plot <- df %>%
  mutate(LD1 = lda_proj[,1])

# ------------------------------------------------
# Gráfico das densidades na direção discriminante
# ------------------------------------------------
ggplot(dados_plot, aes(x = LD1, fill = classe, color = classe)) +
  geom_density(alpha = 0.35, linewidth = 1.2) +
  geom_vline(
    data = dados_plot %>% group_by(classe) %>% summarise(m = mean(LD1)),
    aes(xintercept = m, color = classe),
    linetype = "dashed",
    linewidth = 1.1
  ) +
  labs(
    title = "Projeção na Função Discriminante Linear (LD1)",
    x = "Score discriminante (LD1)",
    y = "Densidade"
  ) +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(face = "bold"),
    legend.position = "top"
  )
```



A projeção das observações na função discriminante linear (LD1) evidencia uma clara separação entre os grupos `Regular` e `Suspeito` ao longo do eixo discriminante. As distribuições de densidade mostram que os indivíduos classificados como regulares concentram-se em valores negativos do escore LD1, enquanto os indivíduos suspeitos apresentam valores substancialmente mais elevados.

A distância entre as médias dos grupos (indicadas pelas linhas tracejadas) sugere forte capacidade discriminante da combinação linear estimada, indicando que o eixo LD1 captura de forma eficaz o contraste multivariado entre padrões comportamentais e operacionais das duas populações. A sobreposição entre as densidades é relativamente pequena, o que indica baixo erro esperado sob uma fronteira de decisão linear.

Além disso, observa-se que o grupo `Suspeito` apresenta maior dispersão ao longo do eixo discriminante, refletindo maior variabilidade estrutural, fato consistente com as diferenças previamente observadas nas matrizes de covariância e no teste de Box. Esse padrão reforça a interpretação de que, embora a separação linear seja forte, a estrutura dos dados pode beneficiar-se de modelos mais flexíveis, como a análise discriminante quadrática, capazes de capturar possíveis assimetrias na dispersão dos grupos.

5.1.1.1 Avaliação do Modelo Linear

```
pred_lda_resub <- predict(fit_lda)$class

conf_lda_resub <- confusionMatrix(pred_lda_resub, df$classe, positive = 'Suspeito')
conf_lda_resub
```

```
Confusion Matrix and Statistics
```

```
          Reference
Prediction Regular Suspeito
  Regular      309       40
  Suspeito      11       40

               Accuracy : 0.8725
                 95% CI : (0.8358, 0.9036)
    No Information Rate : 0.8
    P-Value [Acc > NIR] : 9.165e-05

                  Kappa : 0.5389

 Mcnemar's Test P-Value : 8.826e-05

            Sensitivity : 0.5000
            Specificity : 0.9656
         Pos Pred Value : 0.7843
         Neg Pred Value : 0.8854
             Prevalence : 0.2000
         Detection Rate : 0.1000
   Detection Prevalence : 0.1275
      Balanced Accuracy : 0.7328

       'Positive' Class : Suspeito
```

5.1.2 Análise Discriminante Quadrática (QDA)

```
fit_qda
```

```
Call:
qda(classe ~ ., data = df_scaled)

Prior probabilities of groups:
 Regular Suspeito
     0.8      0.2

Group means:
         valor_reparo dias_para_reportar qt_sinistros_12m ratio_reparo_vs_fipe
Regular    -0.1990037         -0.1289575       -0.1269541           -0.1936770
Suspeito    0.7960148          0.5158298        0.5078163            0.7747081
         dist_evento_resid mudancas_contato_6m
Regular         -0.1227236          -0.1273873
Suspeito         0.4908943           0.5095490
```

A análise discriminante quadrática utiliza as mesmas médias por grupo, porém permite matrizes de covariância específicas para cada classe. Considerando que o teste de Box indicou rejeição da hipótese de homogeneidade das covariâncias e que o grupo `Suspeito` apresenta maior variabilidade estrutural, o modelo QDA é teoricamente mais adequado para representar a separação entre as populações.

Enquanto o LDA produz uma fronteira de decisão linear, o QDA permite superfícies de decisão curvas, capazes de capturar diferenças na dispersão dos grupos. Dado o padrão observado nas matrizes de correlação

e covariância, especialmente a maior intensidade das associações no grupo `Suspeito`, espera-se que o QDA apresente melhor desempenho preditivo em validações fora da amostra.

5.1.2.1 Avaliação do Modelo Quadrático

```
pred_qda_resub <- predict(fit_qda)$class

conf_qda_resub <- confusionMatrix(pred_qda_resub, df$classe, positive = 'Suspeito')
conf_qda_resub
```

```
Confusion Matrix and Statistics

          Reference
Prediction Regular Suspeito
  Regular      302       38
  Suspeito      18       42

               Accuracy : 0.86
                 95% CI : (0.8221, 0.8925)
    No Information Rate : 0.8
    P-Value [Acc > NIR] : 0.001145

                  Kappa : 0.5172

 Mcnemar's Test P-Value : 0.011118

            Sensitivity : 0.5250
            Specificity : 0.9437
         Pos Pred Value : 0.7000
         Neg Pred Value : 0.8882
             Prevalence : 0.2000
         Detection Rate : 0.1050
   Detection Prevalence : 0.1500
      Balanced Accuracy : 0.7344

       'Positive' Class : Suspeito
```

5.2 Método da substituição com Divisão Amostral

```
set.seed(12345)

train_id <- createDataPartition(df$classe, p = 0.7, list = FALSE)

train <- df[train_id, ]
test  <- df[-train_id, ]
```