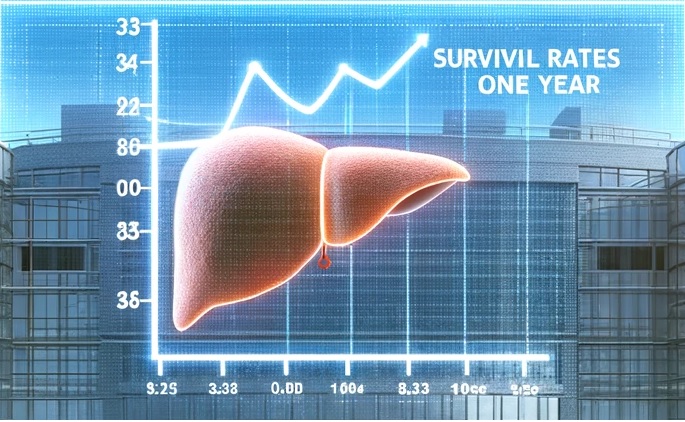
#### **Health Analytics**



#### **1.1 - Problema de Negócio**

O objetivo deste projeto é tentar prever o tempo de sobrevivência de um paciente um ano após o transplante de fígado. Usaremos dados reais disponibilizados publicamente. Há normalmente uma grande fila de pacientes esperando em um determinado momento para receber um transplante de fígado, pois não há outra cura para o estágio final de doença hepática. Nosso estudo visa ajudar os pacientes a compreender melhor suas chances de sobrevivência após um ano do transplante. As informações do nosso estudo mostram que somos capazes de prever o tempo de sobrevivência com razoável precisão para diferentes intervalos de tempo. Nosso modelo pode auxiliar os pacientes no processo de tomada de decisão (como financeiro por exemplo), enquanto eles esperam na fila para fazer um transplante.

Usando dados do Registro Científico de Recipientes de Transplante (Scientic Registry of Transplant Recipients - SRTR), trabalharemos com variáveis quantitativas e qualitativas criando modelos preditivos usando o tempo de sobrevivência do paciente como nossa variável de resposta (ou seja, teremos um problema de regressão, cujo objetivo é prever um valor numérico).

Usaremos duas técnicas de regressão: regressão linear e rede neural. Ambos os modelos previram o tempo de sobrevida do paciente por 1 a 3 anos após o transplante, usando um conjunto de variáveis mais significativas. As variáveis foram escolhidas após uma pesquisa sobre quais são os fatores mais importantes no processo de transplante de fígado.

Nossos resultados nos dão a certeza de que podemos prever com razoável precisão o tempo de sobrevivência um ano após os pacientes receberem um transplante de fígado. Acreditamos que isso dará aos pacientes uma ideia melhor de suas mudanças de sobrevida por um período prolongado de tempo após o transplante. Usando as informações desses modelos, os pacientes podem tomar uma decisão mais informada quando se trata de transplante de órgão com base em sua condição médica.

Os dados foram extraídos do SRTR Database e modificados para que pudessem ser executados com minha capacidade computacional. O Site oficial dos dados: <https://www.srtr.org/about-the-data/the-srtr-database/>. Os dados originais devem ser solicitados no link abaixo e a aprovação pode levar até 30 dias, sendo o objetivo estudo e pesquisa: <https://www.srtr.org/requesting-srtr-data/data-requests/>

#### **1.2 - Dicionário de Dados**

1. **DAYSWAIT\_CHRON**: Número de dias de espera crônicos, possivelmente referindo-se ao tempo que um paciente aguardou por um transplante.
2. **PSTATUS**: Status pós-transplante do paciente, onde ‘0’ indica sem complicações e ‘1’ indica com complicações ou outro estado crítico.
3. **FINAL\_MELD\_SCORE**: Pontuação MELD final (Model for End-Stage Liver Disease) é um número que varia de 6 a 40, com base em testes de laboratório. Ele classifica o grau de doença, que mostra o quanto você precisa de um transplante de fígado. Quanto maior o número, mais urgente é o seu caso.
4. **PTIME**: Tempo pós-transplante, possivelmente medido em dias.
5. **TX\_DATE**: Data do transplante.
6. **PX\_STAT**: Status do paciente, ‘A’ pode representar um estado específico.
7. **PX\_STAT\_DATE**: Data da última atualização do status do paciente.
8. **AGE**: Idade do paciente no momento do transplante.
9. **ABO**: Tipo sanguíneo do paciente.
10. **GENDER**: Gênero do paciente, onde ‘1’ pode representar masculino e ‘0’ feminino.
11. **WGT\_KG\_TCR**: Peso do paciente na altura do cadastro para transplante, em quilogramas.
12. **HGT\_CM\_TCR**: Altura do paciente na altura do cadastro para transplante, em centímetros.
13. **BMI\_TCR**: Índice de Massa Corporal (BMI) do paciente na altura do cadastro para transplante.
14. **DIAB**: Indica se o paciente tem diabetes, com diferentes valores representando diferentes estados ou tipos de diabetes.
15. **INIT\_AGE**: Idade inicial do paciente quando do cadastro para transplante.
16. **ETHCAT**: Categoria étnica do paciente.
17. **REGION**: Região geográfica onde o transplante foi realizado.
18. **PERM\_STATE**: Estado permanente de residência do paciente.
19. **TX\_Year**: Ano em que o transplante foi realizado.
20. **TX\_PROCEDUR\_TY**: Tipo de procedimento de transplante realizado.
21. **MED\_COND\_TRR**: Condição médica do paciente no registro do transplante.
22. **PREV\_TX**: Indica se o paciente já havia recebido um transplante anteriormente (‘N’ para não, ‘Y’ para sim).
23. **AGE\_DON**: Idade do doador.
24. **GENDER\_DON**: Gênero do doador.
25. **HGT\_CM\_DON\_CALC**: Altura calculada do doador, em centímetros.
26. **WGT\_KG\_DON\_CALC**: Peso calculado do doador, em quilogramas.
27. **BMI\_DON\_CALC**: Índice de Massa Corporal calculado do doador.
28. **COD\_CAD\_DON**: Código de causa de morte do doador.
29. **ETHCAT\_DON**: Categoria étnica do doador.
30. **HOME\_STATE\_DON**: Estado de residência do doador.
31. **DIABETES\_DON**: Indica se o doador tinha diabetes (‘Y’ para sim, ‘N’ para não).
32. **HIST\_HYPERTENS\_DON**: Indica se o doador tinha histórico de hipertensão (‘Y’ para sim, ‘N’ para não).
33. **HIST\_IV\_DRUG\_OLD\_DON**: Indica histórico de uso de drogas intravenosas pelo doador.
34. **ABO\_DON**: Tipo sanguíneo do doador.
35. **HIST\_CANCER\_DON**: Indica se o doador tinha histórico de câncer (‘Y’ para sim, ‘N’ para não).
36. **ALCOHOL\_HEAVY\_DON**: Indica se o doador consumia álcool de forma pesada (‘Y’ para sim, ‘N’ para não).
37. **ABO\_MAT**: Compatibilidade do tipo sanguíneo ABO entre doador e receptor.
38. **COLD\_ISCH**: Tempo de isquemia fria, o tempo que o órgão doado permanece preservado em baixas temperaturas durante o transporte até o transplante.
39. **MALIG**: Estado de malignidade, N-não, U-undefined, Y- yes .
40. **HGT\_CM\_CALC**: Altura calculada do paciente, em centímetros. Isso pode ser uma medida registrada ou uma estimativa usada em análises.
41. **WGT\_KG\_CALC**: Peso calculado do paciente, em quilogramas. Assim como a altura, pode ser uma medida atualizada ou uma estimativa.
42. **BMI\_CALC**: Índice de Massa Corporal (BMI) calculado do paciente, uma medida de corporalidade baseada na altura e peso calculados.
43. **TX\_MELD**: Indica se a pontuação MELD foi usada para o transplante. “No” sugere que não foi usado, talvez indicando outro critério ou método para avaliação da prioridade do transplante.
44. **LISTYR**: Ano em que o paciente foi listado para o transplante.
45. **LiverSize**: Tamanho do fígado do paciente, possivelmente medido em centímetros cúbicos ou outra unidade volumétrica, relevante para avaliar a compatibilidade do órgão para transplante.
46. **LiverSizeDon**: Tamanho do fígado do doador, similarmente medido para avaliações de compatibilidade do transplante.

#### **1.3 – Bibliotecas**

library(dplyr)

library(ggcorrplot)

library(forecast)# para calcular a métrica de acurácia

library(nnet) # função para tratamentos dos dados

library(neuralnet) # para criar um modelo de rede neural

library(ggplot2)

library(caret)

library(glmnet)

library(car)

**1.4 - Carregando os dados**

dados <- read.csv("dados/dataset.csv", header =  
TRUE, na.strings = c(""))   
  
dim(dados)

## [1] 79100 46

#### **1.5 - Análise Exploratória, Limpeza, Transformação e Manipulação de Dados**

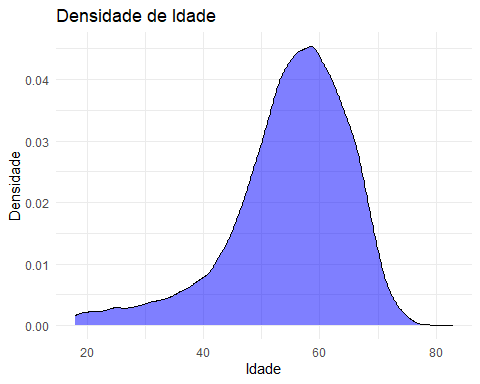
View(dados)

# Tipos dos dados  
str(dados)

## 'data.frame': 79100 obs. of 46 variables:  
## $ DAYSWAIT\_CHRON : int 7 5 10 9 2 6 4 9 1 11 ...  
## $ PSTATUS : int 0 0 0 0 0 0 0 1 1 0 ...  
## $ FINAL\_MELD\_SCORE : int 39 19 22 35 35 19 35 14 36 23 ...  
## $ PTIME : int 51 6 6 27 54 10 51 0 3 6 ...  
## $ TX\_DATE : chr "12/24/2018" "12/23/2018" "12/28/2018" "12/27/2018" ...  
## $ PX\_STAT : chr "A" "A" "A" "A" ...  
## $ PX\_STAT\_DATE : chr "2/13/2019" "12/29/2018" "1/3/2019" "1/23/2019" ...  
## $ AGE : int 30 63 48 54 71 62 62 56 28 66 ...  
## $ ABO : chr "O" "A" "B" "O" ...  
## $ GENDER : int 1 0 0 1 1 1 1 0 1 0 ...  
## $ WGT\_KG\_TCR : num 56.2 81.9 78.9 63.5 75.8 ...  
## $ HGT\_CM\_TCR : num 163 178 181 155 163 ...  
## $ BMI\_TCR : num 21.3 25.9 24.1 26.4 28.7 ...  
## $ DIAB : int 1 1 1 1 1 3 1 1 1 1 ...  
## $ INIT\_AGE : int 30 63 48 54 71 62 62 56 28 66 ...  
## $ ETHCAT : int 1 1 1 1 1 1 1 4 2 1 ...  
## $ REGION : int 2 3 10 4 3 3 5 3 3 11 ...  
## $ PERM\_STATE : chr "MD" "GA" "OH" "TX" ...  
## $ TX\_Year : int 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 ...  
## $ TX\_PROCEDUR\_TY : int 701 701 701 701 701 701 701 701 701 701 ...  
## $ MED\_COND\_TRR : int 1 3 3 2 3 1 1 3 1 3 ...  
## $ PREV\_TX : chr "N" "N" "N" "N" ...  
## $ AGE\_DON : int 24 34 42 48 37 38 23 47 18 35 ...  
## $ GENDER\_DON : int 1 0 1 1 1 0 0 0 0 0 ...  
## $ HGT\_CM\_DON\_CALC : num 173 183 173 157 173 165 175 163 173 178 ...  
## $ WGT\_KG\_DON\_CALC : num 75 90 107 93 81.6 55.7 93 49.9 68 67.7 ...  
## $ BMI\_DON\_CALC : num 25.1 26.9 35.8 37.7 27.3 ...  
## $ COD\_CAD\_DON : int 3 3 1 2 3 3 1 2 3 1 ...  
## $ ETHCAT\_DON : int 2 1 1 1 1 1 4 4 4 1 ...  
## $ HOME\_STATE\_DON : chr "PA" "GA" "NY" "TX" ...  
## $ DIABETES\_DON : chr "N" "N" "Y" "N" ...  
## $ HIST\_HYPERTENS\_DON : chr "N" "N" "Y" "Y" ...  
## $ HIST\_IV\_DRUG\_OLD\_DON: chr NA NA NA NA ...  
## $ ABO\_DON : chr "O" "A" "B" "O" ...  
## $ HIST\_CANCER\_DON : chr "N" "N" "N" "N" ...  
## $ ALCOHOL\_HEAVY\_DON : chr "N" "N" "N" "N" ...  
## $ ABO\_MAT : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ COLD\_ISCH : num 4.3 3.48 4.95 3.62 7.5 5.33 9.14 5.25 4.6 5.6 ...  
## $ MALIG : chr "U" "U" "U" "U" ...  
## $ HGT\_CM\_CALC : num 163 178 183 155 163 ...  
## $ WGT\_KG\_CALC : num 45 85 76.2 61.1 70.8 ...  
## $ BMI\_CALC : num 17 26.9 22.8 25.5 26.8 47.9 19.5 32.1 33.7 24.9 ...  
## $ TX\_MELD : chr "No" "No" "No" "No" ...  
## $ LISTYR : int 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 ...  
## $ LiverSize : num 1722 1935 1987 1669 1605 ...  
## $ LiverSizeDon : num 2277 2387 2555 2255 2215 ...

#### **Explorando os dados das variáveis numéricas**

# Idade do paciente  
plot\_gg <- ggplot(dados, aes(x=AGE)) +  
 geom\_density(fill="blue", alpha=0.5) +  
 labs(title="Densidade de Idade", x="Idade", y="Densidade") +  
 theme\_minimal()  
# Exibe o gráfico no RStudio  
print(plot\_gg)

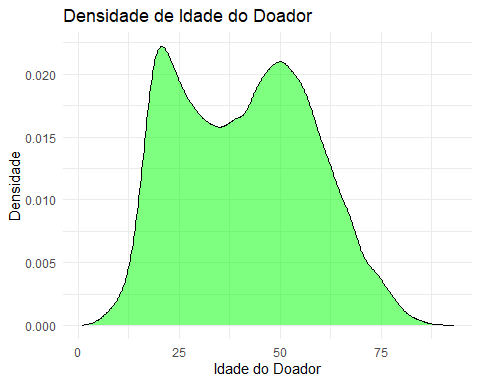


# Salva o gráfico no arquivo usando ggsave  
ggsave("densidade\_idade\_ggplot.png", plot = plot\_gg, width=10, height=8, units="cm")

Análises preliminares:

1. **Forma da Distribuição**: A curva é unimodal, indicando uma única idade mais comum entre os pacientes que receberam transplantes. Não há evidências de múltiplos picos, o que sugeriria grupos distintos de idades.
2. **Localização do Pico Central**: O pico da curva parece estar em torno dos 50-60 anos, indicando que essa é a idade mais comum dos pacientes no momento do transplante.
3. **Assimetria**: O gráfico é ligeiramente assimétrico, com uma cauda mais longa se estendendo para as idades mais jovens e uma queda mais acentuada à medida que as idades aumentam após o pico. Isso pode sugerir que pacientes mais jovens são menos frequentemente submetidos a transplantes em comparação com aqueles na faixa etária de pico.
4. **Amplitude de Idade**: O gráfico mostra que há uma ampla gama de idades entre os pacientes que recebem transplantes, desde jovens adultos até idosos. No entanto, a presença de pacientes muito jovens ou muito velhos parece ser menos comum.

# Idade do doador  
plot\_gg <- ggplot(dados, aes(x=AGE\_DON)) +  
 geom\_density(fill="green", alpha=0.5) +  
 labs(title="Densidade de Idade do Doador", x="Idade do Doador", y="Densidade") +  
 theme\_minimal()  
  
# Exibe o gráfico no RStudio  
print(plot\_gg)



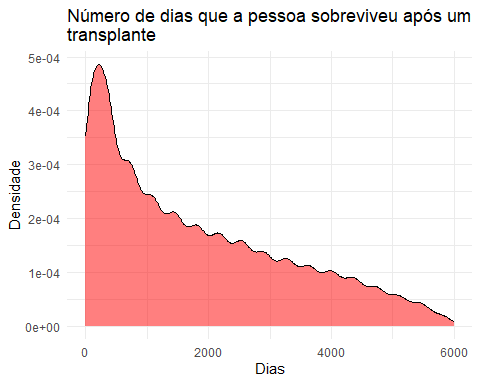
# Salva o gráfico no arquivo usando ggsave  
ggsave("densidade\_idade\_doador.png", plot = plot\_gg, width=10, height=8, units="cm")

Análises preliminares:

1. **Forma da Distribuição**: A distribuição é bimodal, o que significa que existem dois grupos de idades predominantes entre os doadores. Isso pode sugerir que existem dois perfis demográficos distintos entre os doadores.
2. **Picos da Distribuição**: Há dois picos visíveis. O primeiro pico é em torno dos 20 anos, e o segundo, menor, está por volta dos 50 anos. Isso pode indicar que a maioria dos doadores tende a ser mais jovem, com um segundo grupo significativo de meia-idade.
3. **Variação de Idade**: O gráfico cobre uma gama de idades de 0 até aproximadamente 80 anos, mostrando que os doadores vêm de todas as faixas etárias, mas com menos frequência nos extremos da distribuição.
4. **Assimetria**: A distribuição apresenta assimetria positiva (ou à direita), com uma cauda mais longa estendendo-se em direção a idades mais avançadas. Isso indica que, embora existam doadores mais velhos, eles são menos frequentes.
5. **Interpretação Clínica**: A bimodalidade pode sugerir que há um grupo de doadores mais jovens, que talvez sejam doadores vivos ou tenham sofrido mortes prematuras, e um grupo de doadores mais velhos, que talvez sejam casos de morte natural ou condições crônicas. A presença de doadores mais jovens pode estar relacionada a causas de morte acidentais ou traumáticas.
6. **Considerações para a Política de Doação**: Informações como essas podem ser importantes para desenvolver estratégias de educação e recrutamento de doadores, considerando as idades que mais frequentemente estão presentes na população doadora.

# Número de dias que a pessoa sobreviveu após um transplante  
plot\_gg <- ggplot(dados, aes(x=PTIME)) + geom\_density(fill="red",  
alpha=0.5) + labs(title="Número de dias que a pessoa sobreviveu após um  
transplante", x="Dias", y="Densidade") + theme\_minimal()  
  
# Exibe o gráfico no RStudio  
print(plot\_gg)

## Warning: Removed 4 rows containing non-finite values (`stat\_density()`).



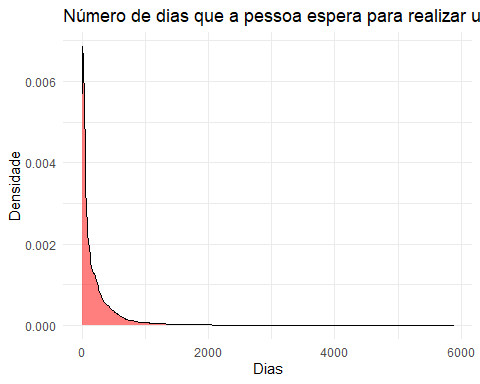
# Salva o gráfico no arquivo usando ggsave  
ggsave("numero\_dias\_apos\_transpl.png", plot = plot\_gg, width=10,  
height=8, units="cm")

## Warning: Removed 4 rows containing non-finite values (`stat\_density()`).

Análises preliminares:

1. **Pico Inicial Acentuado**: Existe um pico muito acentuado perto do zero. Isso sugere que uma quantidade relativamente grande de indivíduos não sobrevive muito tempo após o transplante.
2. **Declínio Rápido**: Após o pico inicial, há uma queda rápida e constante na frequência conforme aumentam os dias de sobrevivência. Isso indica que menos pessoas sobrevivem por períodos mais longos após o transplante.
3. **Cauda Longa**: A distribuição tem uma cauda longa que se estende para a direita, indicando que, embora em menor número, há indivíduos que sobrevivem por muitos dias após o transplante.
4. **Possíveis Complicações Precoce**: A grande frequência de sobrevivência por um período curto pode refletir complicações pós-operatórias imediatas ou rejeição precoce do órgão transplantado.
5. **Sobrevivência a Longo Prazo**: Mesmo com a frequência diminuindo, a presença de uma cauda longa indica que alguns pacientes conseguem uma sobrevivência a longo prazo.
6. **Análise Quantitativa**: Seria importante calcular métricas como a mediana e a média de sobrevivência para entender melhor o tempo típico de sobrevivência. Outra análise relevante seria a sobrevida cumulativa ao longo do tempo.
7. **Interpretação Clínica e Qualidade de Vida**: A análise detalhada dessa distribuição pode ajudar a entender e a melhorar os protocolos clínicos de acompanhamento e tratamento pós-transplante para aumentar a sobrevivência e a qualidade de vida dos pacientes.

# Número de dias que a pessoa espera para realizar um transplante  
  
plot\_gg <- ggplot(dados, aes(x=DAYSWAIT\_CHRON)) + geom\_density(fill="red",  
alpha=0.5) + labs(title="Número de dias que a pessoa espera para realizar um transplante", x="Dias", y="Densidade") + theme\_minimal()  
  
# Exibe o gráfico no RStudio  
  
print(plot\_gg)



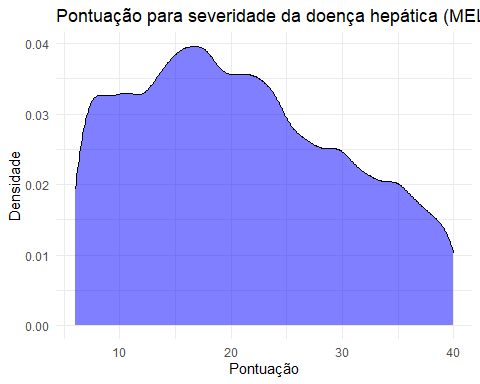
# Salva o gráfico no arquivo usando ggsave  
  
ggsave("numero\_dias\_espera\_transpl.png", plot = plot\_gg, width=10,  
height=8, units="cm")

Análises Preliminares:

1. **Pico Próximo a Zero**: Existe um pico muito proeminente perto de zero, o que sugere que uma grande proporção de pacientes recebe um transplante logo após serem listados.

1. **Declínio Acentuado**: O gráfico mostra um declínio acentuado na densidade conforme o número de dias aumenta, indicando que o número de pacientes que esperam por um período mais longo diminui rapidamente.
2. **Cauda Longa**: A cauda longa da distribuição sugere que, embora raro, há pacientes que esperam um tempo muito longo para um transplante. Esta cauda se estende até aproximadamente 6000 dias.
3. **Grande Maioria Espera Pouco Tempo**: A área sob a curva é mais concentrada perto do início do eixo dos dias, o que significa que a grande maioria dos pacientes espera relativamente pouco tempo.
4. **Possíveis Melhorias no Sistema de Alocação**: Se o objetivo é minimizar o tempo de espera, o gráfico sugere que o sistema de alocação está funcionando bem para a maioria dos pacientes, mas ainda há espaço para melhorias para aqueles que acabam esperando mais tempo.
5. **Considerações Adicionais**: Seria importante entender os fatores que contribuem para os tempos de espera mais longos para alguns pacientes. Isso pode incluir tipo de sangue, compatibilidade do órgão, gravidade da condição médica, ou outros critérios utilizados para alocar órgãos para transplante.

# Pontuação para severidade da doença hepática (MELD)  
plot\_gg <- ggplot(dados, aes(x=FINAL\_MELD\_SCORE)) + geom\_density(fill="blue",  
alpha=0.5) + labs(title="Pontuação para severidade da doença hepática (MELD)", x="Pontuação", y="Densidade") + theme\_minimal()  
  
# Exibe o gráfico no RStudio  
print(plot\_gg)



# Salva o gráfico no arquivo usando ggsave  
ggsave("pontuação\_MELD.png", plot = plot\_gg, width=10,  
height=8, units="cm")

#### Explorando as variáveis categóricas

library(skimr)

## Warning: package 'skimr' was built under R version 4.3.3

# Resumo rápido para EDA  
skim(dados)

Data summary

|  |  |
| --- | --- |
| Name | dados |
| Number of rows | 79100 |
| Number of columns | 46 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 15 |
| numeric | 31 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| TX\_DATE | 0 | 1.00 | 8 | 10 | 0 | 6139 | 0 |
| PX\_STAT | 0 | 1.00 | 1 | 1 | 0 | 2 | 0 |
| PX\_STAT\_DATE | 0 | 1.00 | 1 | 10 | 0 | 5774 | 0 |
| ABO | 0 | 1.00 | 1 | 3 | 0 | 8 | 0 |
| PERM\_STATE | 0 | 1.00 | 2 | 2 | 0 | 56 | 0 |
| PREV\_TX | 0 | 1.00 | 1 | 1 | 0 | 2 | 0 |
| HOME\_STATE\_DON | 390 | 1.00 | 2 | 2 | 0 | 57 | 0 |
| DIABETES\_DON | 2 | 1.00 | 1 | 1 | 0 | 3 | 0 |
| HIST\_HYPERTENS\_DON | 2 | 1.00 | 1 | 1 | 0 | 3 | 0 |
| HIST\_IV\_DRUG\_OLD\_DON | 73119 | 0.08 | 1 | 1 | 0 | 3 | 0 |
| ABO\_DON | 0 | 1.00 | 1 | 3 | 0 | 8 | 0 |
| HIST\_CANCER\_DON | 2 | 1.00 | 1 | 1 | 0 | 3 | 0 |
| ALCOHOL\_HEAVY\_DON | 5717 | 0.93 | 1 | 1 | 0 | 3 | 0 |
| MALIG | 0 | 1.00 | 1 | 1 | 0 | 3 | 0 |
| TX\_MELD | 0 | 1.00 | 2 | 3 | 0 | 2 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| DAYSWAIT\_CHRON | 0 | 1.00 | 228.53 | 400.51 | 0.00 | 19.00 | 87.00 | 266.00 | 5886.00 | ▇▁▁▁▁ |
| PSTATUS | 0 | 1.00 | 0.30 | 0.46 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 | ▇▁▁▁▃ |
| FINAL\_MELD\_SCORE | 0 | 1.00 | 20.87 | 9.20 | 6.00 | 13.00 | 20.00 | 28.00 | 40.00 | ▆▇▇▅▃ |
| PTIME | 4 | 1.00 | 1792.61 | 1538.56 | 0.00 | 390.00 | 1431.00 | 2889.00 | 6001.00 | ▇▃▃▂▁ |
| AGE | 0 | 1.00 | 54.97 | 10.24 | 18.00 | 50.00 | 56.00 | 62.00 | 83.00 | ▁▂▇▇▁ |
| GENDER | 0 | 1.00 | 0.33 | 0.47 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 | ▇▁▁▁▃ |
| WGT\_KG\_TCR | 103 | 1.00 | 85.34 | 19.72 | 9.53 | 71.50 | 83.55 | 97.52 | 263.00 | ▁▇▁▁▁ |
| HGT\_CM\_TCR | 166 | 1.00 | 172.13 | 10.22 | 6.30 | 165.10 | 172.72 | 180.30 | 225.00 | ▁▁▁▇▂ |
| BMI\_TCR | 185 | 1.00 | 29.09 | 102.72 | 2.85 | 24.59 | 28.03 | 32.19 | 28798.19 | ▇▁▁▁▁ |
| DIAB | 11 | 1.00 | 12.03 | 101.42 | 1.00 | 1.00 | 1.00 | 3.00 | 998.00 | ▇▁▁▁▁ |
| INIT\_AGE | 0 | 1.00 | 54.35 | 10.22 | 13.00 | 49.00 | 56.00 | 61.00 | 83.00 | ▁▁▆▇▁ |
| ETHCAT | 0 | 1.00 | 1.73 | 1.39 | 1.00 | 1.00 | 1.00 | 2.00 | 9.00 | ▇▁▁▁▁ |
| REGION | 0 | 1.00 | 5.82 | 3.16 | 1.00 | 3.00 | 5.00 | 9.00 | 11.00 | ▇▅▃▃▅ |
| TX\_Year | 0 | 1.00 | 2011.17 | 4.70 | 2001.00 | 2007.00 | 2011.00 | 2015.00 | 2018.00 | ▂▅▆▅▇ |
| TX\_PROCEDUR\_TY | 0 | 1.00 | 701.03 | 0.23 | 701.00 | 701.00 | 701.00 | 701.00 | 704.00 | ▇▁▁▁▁ |
| MED\_COND\_TRR | 5 | 1.00 | 2.62 | 0.67 | 1.00 | 2.00 | 3.00 | 3.00 | 3.00 | ▁▁▂▁▇ |
| AGE\_DON | 0 | 1.00 | 41.14 | 16.71 | 1.00 | 26.00 | 42.00 | 54.00 | 93.00 | ▂▇▇▅▁ |
| GENDER\_DON | 0 | 1.00 | 0.40 | 0.49 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 | ▇▁▁▁▆ |
| HGT\_CM\_DON\_CALC | 3 | 1.00 | 171.49 | 10.85 | 86.00 | 165.00 | 172.70 | 180.00 | 213.40 | ▁▁▂▇▁ |
| WGT\_KG\_DON\_CALC | 43 | 1.00 | 80.88 | 20.41 | 16.00 | 67.10 | 78.80 | 92.00 | 200.00 | ▁▇▃▁▁ |
| BMI\_DON\_CALC | 107 | 1.00 | 27.43 | 6.34 | 10.25 | 23.09 | 26.39 | 30.54 | 74.36 | ▃▇▁▁▁ |
| COD\_CAD\_DON | 2 | 1.00 | 22.30 | 140.52 | 1.00 | 1.00 | 2.00 | 3.00 | 999.00 | ▇▁▁▁▁ |
| ETHCAT\_DON | 10 | 1.00 | 1.74 | 1.37 | 1.00 | 1.00 | 1.00 | 2.00 | 9.00 | ▇▁▁▁▁ |
| ABO\_MAT | 0 | 1.00 | 1.07 | 0.29 | 1.00 | 1.00 | 1.00 | 1.00 | 3.00 | ▇▁▁▁▁ |
| COLD\_ISCH | 2384 | 0.97 | 6.64 | 2.93 | 0.00 | 4.92 | 6.16 | 8.00 | 48.00 | ▇▁▁▁▁ |
| HGT\_CM\_CALC | 18 | 1.00 | 172.13 | 10.21 | 123.40 | 165.10 | 172.70 | 180.30 | 213.40 | ▁▂▇▅▁ |
| WGT\_KG\_CALC | 13 | 1.00 | 84.13 | 19.52 | 33.60 | 70.00 | 82.20 | 96.20 | 202.30 | ▂▇▂▁▁ |
| BMI\_CALC | 47 | 1.00 | 28.30 | 5.71 | 15.00 | 24.20 | 27.60 | 31.80 | 72.90 | ▆▇▁▁▁ |
| LISTYR | 0 | 1.00 | 2010.55 | 4.70 | 2001.00 | 2007.00 | 2011.00 | 2015.00 | 2018.00 | ▅▅▇▆▇ |
| LiverSize | 0 | 1.00 | 2023.91 | 332.42 | 631.26 | 1790.75 | 1995.83 | 2227.74 | 3977.01 | ▁▇▇▁▁ |
| LiverSizeDon | 0 | 1.00 | 2149.35 | 368.16 | 579.82 | 1903.45 | 2141.24 | 2371.93 | 4224.40 | ▁▅▇▁▁ |

A variável abaixo está um pouco confusa em relação aos seus dados.Se não tivermos certeza do tipo da variável descarte-a

dados$DIAB <- as.factor(dados$DIAB)  
table(dados$DIAB)

##   
## 1 2 3 4 5 998   
## 57017 1520 16476 309 2939 828

* A variável abaixo indica se a pessoa sobrevive ou não ao transplante.
* 0 indica a classe negativa e indica que a pessoa sobrevive.
* 1 indica a classe positiva indica que a pessoa não sobrevive.

dados$PSTATUS <- as.factor(dados$PSTATUS)  
table(dados$PSTATUS)

##   
## 0 1   
## 55634 23466

* Gênero do recebedor

dados$GENDER <- as.factor(dados$GENDER)  
table(dados$GENDER)

##   
## 0 1   
## 53312 25788

* Gênero do doador

dados$GENDER\_DON <- as.factor(dados$GENDER\_DON)  
table(dados$GENDER\_DON)

##   
## 0 1   
## 47310 31790

* Fonte de dados da região de onde foram coletados os dados.

##   
## 1 2 3 4 5 6 7 8 9 10 11   
## 2802 9435 14070 7436 9962 2505 6503 5458 4745 7656 8528

* A variável abaixo indica o ano em que os dados foram registrados.
* Os anos de 2001 e 2002 apresentam valores bem pequenos em relação aos demais.
* A decisão será , mais adiante, não utilizar os registros deste anos.

dados$TX\_Year <- as.factor(dados$TX\_Year)  
table(dados$TX\_Year)

##   
## 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016   
## 1 1456 2948 3717 4062 4475 4501 4459 4641 4583 4745 4751 4898 5128 5430 6239   
## 2017 2018   
## 6519 6547

* A variável abaixo indica se havia ou não algum tipo de tumor no recebedor.
* N-não, U-undefined, Y- yes

dados$MALIG <- as.factor(dados$MALIG)  
table(dados$MALIG)

##   
## N U Y   
## 42828 21290 14982

* histórico de cancer do doador
* N-não, U-undefined, Y- yes

dados$HIST\_CANCER\_DON <- as.factor(dados$HIST\_CANCER\_DON)  
table(dados$HIST\_CANCER\_DON)

##   
## N U Y   
## 76040 398 2660

* Considerando apenas os pacientes que sobreviveram ao primeiro ano de cirurgia,pois queremos prever o tempo de sobrevivência após um ano de transplante

dados1 <- dados %>%  
 filter(PTIME > 365) %>%  
 mutate(PTIME = (PTIME - 365))  
  
dim(dados1)

## [1] 61600 46

* Dos pacientes que sobreviveram ao primeiro ano da cirurgia, filtramos os que permaneceram vivos por até três anos depois da cirurgia.

dados2 <- dados1 %>%  
 filter(PTIME <= 1095)  
  
dim(dados2)

## [1] 23348 46

View(dados2)  
str(dados2)

## 'data.frame': 23348 obs. of 46 variables:  
## $ DAYSWAIT\_CHRON : int 97 744 35 112 136 20 49 1 99 3435 ...  
## $ PSTATUS : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...  
## $ FINAL\_MELD\_SCORE : int 22 26 20 26 28 29 21 13 13 29 ...  
## $ PTIME : num 728 685 183 69 359 965 704 575 495 42 ...  
## $ TX\_DATE : chr "6/5/2002" "3/13/2004" "4/9/2002" "6/25/2002" ...  
## $ PX\_STAT : chr "D" "D" "D" "D" ...  
## $ PX\_STAT\_DATE : chr "6/2/2005" "1/27/2007" "10/9/2003" "9/2/2003" ...  
## $ AGE : int 58 58 47 52 49 51 60 45 22 35 ...  
## $ ABO : chr "A" "O" "O" "A" ...  
## $ GENDER : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 2 1 2 2 ...  
## $ WGT\_KG\_TCR : num 87.5 91 79.4 64.9 66 ...  
## $ HGT\_CM\_TCR : num 180 157 163 168 178 ...  
## $ BMI\_TCR : num 26.9 36.7 30 23.1 20.9 ...  
## $ DIAB : Factor w/ 6 levels "1","2","3","4",..: 5 1 1 1 1 1 1 6 1 1 ...  
## $ INIT\_AGE : int 58 56 47 51 49 51 60 45 22 25 ...  
## $ ETHCAT : int 1 4 1 1 1 2 4 1 1 1 ...  
## $ REGION : Factor w/ 11 levels "1","2","3","4",..: 11 5 9 9 5 10 2 4 2 8 ...  
## $ PERM\_STATE : chr "AR" "CA" "NY" "NY" ...  
## $ TX\_Year : Factor w/ 18 levels "2001","2002",..: 2 4 2 2 2 2 2 2 2 11 ...  
## $ TX\_PROCEDUR\_TY : int 701 701 701 701 701 701 701 701 701 701 ...  
## $ MED\_COND\_TRR : int 3 2 3 3 2 3 3 3 3 3 ...  
## $ PREV\_TX : chr "N" "N" "Y" "N" ...  
## $ AGE\_DON : int 43 56 48 17 21 61 13 58 24 28 ...  
## $ GENDER\_DON : Factor w/ 2 levels "0","1": 1 1 2 1 1 2 2 2 2 1 ...  
## $ HGT\_CM\_DON\_CALC : num 178 163 158 175 173 ...  
## $ WGT\_KG\_DON\_CALC : num 96.7 62.3 129 72 68 89.7 59.9 91.2 50 108 ...  
## $ BMI\_DON\_CALC : num 30.6 23.6 52 23.4 22.7 ...  
## $ COD\_CAD\_DON : int 3 2 2 3 3 2 1 2 3 999 ...  
## $ ETHCAT\_DON : int 1 4 2 1 1 1 2 2 1 1 ...  
## $ HOME\_STATE\_DON : chr "TN" "CA" "NY" "NY" ...  
## $ DIABETES\_DON : chr "N" "Y" "Y" "N" ...  
## $ HIST\_HYPERTENS\_DON : chr "N" "Y" "Y" "N" ...  
## $ HIST\_IV\_DRUG\_OLD\_DON: chr "N" "N" "N" "N" ...  
## $ ABO\_DON : chr "A" "O" "O" "A" ...  
## $ HIST\_CANCER\_DON : Factor w/ 3 levels "N","U","Y": 1 1 1 1 1 1 1 1 1 1 ...  
## $ ALCOHOL\_HEAVY\_DON : chr NA NA NA NA ...  
## $ ABO\_MAT : int 1 1 1 1 1 2 1 2 1 1 ...  
## $ COLD\_ISCH : num 7 6 10.3 NA NA 3 11.2 0.1 7 10 ...  
## $ MALIG : Factor w/ 3 levels "N","U","Y": 1 3 1 1 1 1 1 2 3 3 ...  
## $ HGT\_CM\_CALC : num 180 157 165 168 178 ...  
## $ WGT\_KG\_CALC : num 87.7 91 82.6 68.5 62.7 ...  
## $ BMI\_CALC : num 27 36.9 30.3 24.4 19.8 30 27.4 26.6 33.2 20.7 ...  
## $ TX\_MELD : chr "No" "No" "No" "No" ...  
## $ LISTYR : int 2002 2002 2002 2002 2002 2002 2002 2002 2002 2002 ...  
## $ LiverSize : num 2042 2095 2102 1813 1759 ...  
## $ LiverSizeDon : num 2378 1663 2829 2319 2204 ...

**Retirando valores NA**

dim(dados2)

## [1] 23348 46

sum(is.na(dados2))

## [1] 23912

**Transformando variáveis categóricas em fator:**

# Selecionar variáveis do tipo character e transformá-las em factor  
dados3 <- dados2 %>%  
 mutate(across(where(is.character), as.factor))  
summary(dados3)

## DAYSWAIT\_CHRON PSTATUS FINAL\_MELD\_SCORE PTIME   
## Min. : 0.0 0:16335 Min. : 6.00 Min. : 1.0   
## 1st Qu.: 23.0 1: 7013 1st Qu.:13.00 1st Qu.: 252.0   
## Median : 107.0 Median :20.00 Median : 404.0   
## Mean : 255.8 Mean :20.85 Mean : 492.6   
## 3rd Qu.: 302.0 3rd Qu.:28.00 3rd Qu.: 738.0   
## Max. :5425.0 Max. :40.00 Max. :1095.0   
##   
## TX\_DATE PX\_STAT PX\_STAT\_DATE AGE   
## 11/30/2016: 28 A:16447 6/19/2018 : 84 Min. :18.00   
## 8/7/2015 : 28 D: 6901 3/28/2018 : 80 1st Qu.:51.00   
## 12/2/2016 : 26 5/22/2018 : 80 Median :58.00   
## 12/18/2016: 25 7/18/2018 : 79 Mean :55.74   
## 7/27/2015 : 25 8/6/2018 : 76 3rd Qu.:63.00   
## 1/18/2017 : 24 12/17/2018: 75 Max. :81.00   
## (Other) :23192 (Other) :22874   
## ABO GENDER WGT\_KG\_TCR HGT\_CM\_TCR BMI\_TCR   
## O :10315 0:15903 Min. : 24.40 Min. : 22.86 Min. : 9.76   
## A : 8599 1: 7445 1st Qu.: 71.94 1st Qu.:165.10 1st Qu.: 24.69   
## B : 3194 Median : 83.92 Median :172.72 Median : 28.17   
## AB : 1208 Mean : 85.73 Mean :172.20 Mean : 28.87   
## A1 : 23 3rd Qu.: 97.80 3rd Qu.:180.30 3rd Qu.: 32.26   
## A2 : 5 Max. :223.00 Max. :225.00 Max. :1536.34   
## (Other): 4 NA's :24 NA's :33 NA's :39   
## DIAB INIT\_AGE ETHCAT REGION PERM\_STATE   
## 1 :16423 Min. :14.00 Min. :1.000 3 :4181 CA : 2373   
## 2 : 400 1st Qu.:50.00 1st Qu.:1.000 5 :3048 TX : 2092   
## 3 : 5760 Median :57.00 Median :1.000 2 :2917 FL : 1598   
## 4 : 126 Mean :55.03 Mean :1.735 11 :2517 NY : 1267   
## 5 : 528 3rd Qu.:62.00 3rd Qu.:2.000 4 :2363 PA : 1181   
## 998: 111 Max. :81.00 Max. :9.000 10 :2172 IL : 996   
## (Other):6150 (Other):13841   
## TX\_Year TX\_PROCEDUR\_TY MED\_COND\_TRR PREV\_TX AGE\_DON   
## 2016 :5636 Min. :701 Min. :1.000 N:22267 Min. : 2.00   
## 2015 :4781 1st Qu.:701 1st Qu.:2.000 Y: 1081 1st Qu.:27.00   
## 2017 :3271 Median :701 Median :3.000 Median :42.00   
## 2014 :2940 Mean :701 Mean :2.621 Mean :41.56   
## 2013 : 660 3rd Qu.:701 3rd Qu.:3.000 3rd Qu.:54.00   
## 2009 : 609 Max. :704 Max. :3.000 Max. :93.00   
## (Other):5451   
## GENDER\_DON HGT\_CM\_DON\_CALC WGT\_KG\_DON\_CALC BMI\_DON\_CALC COD\_CAD\_DON   
## 0:14102 Min. : 91.0 Min. : 16.00 Min. :10.25 Min. : 1.00   
## 1: 9246 1st Qu.:165.0 1st Qu.: 68.00 1st Qu.:23.32 1st Qu.: 1.00   
## Median :172.7 Median : 79.70 Median :26.72 Median : 2.00   
## Mean :171.4 Mean : 81.88 Mean :27.80 Mean : 22.75   
## 3rd Qu.:180.0 3rd Qu.: 93.20 3rd Qu.:31.06 3rd Qu.: 3.00   
## Max. :209.6 Max. :200.00 Max. :70.50 Max. :999.00   
## NA's :1 NA's :14 NA's :32   
## ETHCAT\_DON HOME\_STATE\_DON DIABETES\_DON HIST\_HYPERTENS\_DON  
## Min. :1.000 TX : 2076 N:20543 N:14982   
## 1st Qu.:1.000 CA : 1965 U: 128 U: 143   
## Median :1.000 FL : 1728 Y: 2677 Y: 8223   
## Mean :1.779 PA : 1334   
## 3rd Qu.:2.000 NY : 1046   
## Max. :9.000 (Other):15119   
## NA's :2 NA's : 80   
## HIST\_IV\_DRUG\_OLD\_DON ABO\_DON HIST\_CANCER\_DON ALCOHOL\_HEAVY\_DON  
## N : 737 O :10889 N:22440 N :18559   
## U : 9 A1 : 4578 U: 124 U : 422   
## Y : 17 A : 3461 Y: 784 Y : 3638   
## NA's:22585 B : 2857 NA's: 729   
## A2 : 784   
## AB : 328   
## (Other): 451   
## ABO\_MAT COLD\_ISCH MALIG HGT\_CM\_CALC WGT\_KG\_CALC   
## Min. :1.000 Min. : 0.100 N: 7144 Min. :132.0 Min. : 34.90   
## 1st Qu.:1.000 1st Qu.: 4.750 U:10984 1st Qu.:165.1 1st Qu.: 70.40   
## Median :1.000 Median : 6.000 Y: 5220 Median :172.7 Median : 82.60   
## Mean :1.074 Mean : 6.365 Mean :172.2 Mean : 84.44   
## 3rd Qu.:1.000 3rd Qu.: 7.600 3rd Qu.:180.3 3rd Qu.: 96.60   
## Max. :3.000 Max. :48.000 Max. :213.4 Max. :181.90   
## NA's :348 NA's :5 NA's :5   
## BMI\_CALC TX\_MELD LISTYR LiverSize LiverSizeDon   
## Min. :15.00 No:23348 Min. :2002 Min. : 695.6 Min. : 708.4   
## 1st Qu.:24.30 1st Qu.:2011 1st Qu.:1784.5 1st Qu.:1910.2   
## Median :27.70 Median :2014 Median :1989.7 Median :2149.8   
## Mean :28.38 Mean :2013 Mean :2019.0 Mean :2160.0   
## 3rd Qu.:31.90 3rd Qu.:2016 3rd Qu.:2221.6 3rd Qu.:2382.9   
## Max. :61.30 Max. :2018 Max. :3952.6 Max. :4102.5   
## NA's :15

Retirando algumas variáveis que não farão parte da análise. HIST\_IV\_DRUG\_OLD\_DON possui quase a totalidade de valores NA

dados3 <- subset(dados3, select = -c(PX\_STAT\_DATE, TX\_MELD, TX\_DATE, HIST\_IV\_DRUG\_OLD\_DON))

Verifcando valores NA

sum(is.na.data.frame(dados3))

## [1] 1327

Retirando valores NA

dados3\_sem\_na <- na.omit(dados3)  
  
dim(dados3\_sem\_na)

## [1] 22218 42

Separando variáveis numéricas de categóricas

dados\_num <- dados3\_sem\_na %>%  
 select\_if(~!is.factor(.))  
  
dados\_fator <- dados3\_sem\_na %>%  
 select\_if(~is.factor(.))  
  
  
dim(dados\_num)

## [1] 22218 25

dim(dados\_fator)

## [1] 22218 17

str(dados\_num)

## 'data.frame': 22218 obs. of 25 variables:  
## $ DAYSWAIT\_CHRON : int 3435 815 1545 2022 1086 1563 1174 1460 1774 1039 ...  
## $ FINAL\_MELD\_SCORE: int 29 17 12 11 17 8 27 12 16 18 ...  
## $ PTIME : num 42 739 54 127 224 547 526 315 18 717 ...  
## $ AGE : int 35 59 48 53 50 57 54 57 71 57 ...  
## $ WGT\_KG\_TCR : num 58.1 84.4 93.9 107 88.5 ...  
## $ HGT\_CM\_TCR : num 163 168 190 190 170 ...  
## $ BMI\_TCR : num 22 30 25.9 29.6 30.5 ...  
## $ INIT\_AGE : int 25 57 44 47 47 52 51 53 66 54 ...  
## $ ETHCAT : int 1 4 1 1 1 5 1 1 1 4 ...  
## $ TX\_PROCEDUR\_TY : int 701 701 701 701 701 701 701 701 701 701 ...  
## $ MED\_COND\_TRR : int 3 3 3 3 3 2 3 2 3 3 ...  
## $ AGE\_DON : int 28 21 23 44 69 15 50 53 21 57 ...  
## $ HGT\_CM\_DON\_CALC : num 175 178 168 168 178 ...  
## $ WGT\_KG\_DON\_CALC : num 108 88 72.2 69.4 64 43 86.2 94 68 58 ...  
## $ BMI\_DON\_CALC : num 35.3 27.8 25.7 24.7 20.2 ...  
## $ COD\_CAD\_DON : int 999 3 3 2 2 1 3 2 3 3 ...  
## $ ETHCAT\_DON : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ ABO\_MAT : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ COLD\_ISCH : num 10 10 9 3.2 9.3 9 8.3 9.1 7 7 ...  
## $ HGT\_CM\_CALC : num 173 178 188 193 170 ...  
## $ WGT\_KG\_CALC : num 61.7 86.2 113.9 101 73.9 ...  
## $ BMI\_CALC : num 20.7 27.3 32.2 27.1 25.5 25 33.2 25.4 20.3 30 ...  
## $ LISTYR : int 2002 2002 2002 2002 2002 2002 2002 2002 2002 2002 ...  
## $ LiverSize : num 1923 2005 2588 2318 1925 ...  
## $ LiverSizeDon : num 2751 2523 2245 1930 1523 ...  
## - attr(\*, "na.action")= 'omit' Named int [1:1130] 1 2 3 4 5 6 7 8 9 11 ...  
## ..- attr(\*, "names")= chr [1:1130] "1" "2" "3" "4" ...

str(dados\_fator)

## 'data.frame': 22218 obs. of 17 variables:  
## $ PSTATUS : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...  
## $ PX\_STAT : Factor w/ 2 levels "A","D": 2 2 2 2 2 2 2 2 2 2 ...  
## $ ABO : Factor w/ 8 levels "A","A1","A1B",..: 8 6 1 6 1 8 2 8 6 7 ...  
## $ GENDER : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 1 2 1 ...  
## $ DIAB : Factor w/ 6 levels "1","2","3","4",..: 1 1 1 1 5 1 1 5 1 1 ...  
## $ REGION : Factor w/ 11 levels "1","2","3","4",..: 8 3 11 5 2 9 3 5 7 4 ...  
## $ PERM\_STATE : Factor w/ 55 levels "AK","AL","AR",..: 6 11 29 5 48 33 10 5 4 46 ...  
## $ TX\_Year : Factor w/ 18 levels "2001","2002",..: 11 4 6 7 5 6 5 6 7 5 ...  
## $ PREV\_TX : Factor w/ 2 levels "N","Y": 1 1 1 1 1 1 1 1 1 1 ...  
## $ GENDER\_DON : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 1 1 2 ...  
## $ HOME\_STATE\_DON : Factor w/ 56 levels "AK","AL","AR",..: 14 11 30 6 49 37 11 36 16 47 ...  
## $ DIABETES\_DON : Factor w/ 3 levels "N","U","Y": 1 1 1 1 1 1 1 1 1 1 ...  
## $ HIST\_HYPERTENS\_DON: Factor w/ 3 levels "N","U","Y": 1 1 1 3 1 1 3 3 1 1 ...  
## $ ABO\_DON : Factor w/ 8 levels "A","A1","A1B",..: 8 6 2 6 4 8 4 8 5 7 ...  
## $ HIST\_CANCER\_DON : Factor w/ 3 levels "N","U","Y": 1 1 1 1 1 1 1 1 1 1 ...  
## $ ALCOHOL\_HEAVY\_DON : Factor w/ 3 levels "N","U","Y": 1 1 1 3 2 1 1 3 1 1 ...  
## $ MALIG : Factor w/ 3 levels "N","U","Y": 3 1 1 1 1 3 1 1 1 1 ...  
## - attr(\*, "na.action")= 'omit' Named int [1:1130] 1 2 3 4 5 6 7 8 9 11 ...  
## ..- attr(\*, "names")= chr [1:1130] "1" "2" "3" "4" ...

View(dados\_num)  
View(dados\_fator)

**Nota**: Para evitar a “armadilha dummy”, que é um problema de multicolinearidade perfeita resultante da inclusão de uma variável dummy para cada nível de uma variável categórica, é essencial deixar de fora uma das categorias dummy de cada variável categórica. A categoria que você deixa de fora atua como a categoria de referência ou base, e os coeficientes das outras categorias dummy são interpretados em relação a essa categoria base.

# Convertendo variáveis categóricas em dummies. O modelo não inclui o intercepto (-1), então uma categoria é automaticamente excluída.  
dados\_fator\_dummies <- model.matrix(~ . - 1, data = dados\_fator)  
  
# Combinando os dados numéricos e os dummies das variáveis categóricas. glmnet faz a padronização automaticamente  
dados\_combinados <- cbind(dados\_num, dados\_fator\_dummies)  
  
#Variável alvo na primeira posição  
dados\_combinados <- dados\_combinados %>%  
 select(PTIME, everything())  
  
#padronizando variáveis numéricas execto variável alvo  
dados\_combinados[, 2:25] <-scale(dados\_combinados[, 2:25])  
dim(dados\_combinados)

## [1] 22218 196

View(dados\_combinados)

**Verificando a multcolinearidade por meio do cálculo do VIF**

O VIF(Fator de inflação de Variância) mede o quanto a variância de um coeficiente estimado é aumentada devido à multicolinearidade. Valores de VIF maiores que 5 ou 10 (dependendo da fonte literária ou da tolerância específica do seu projeto) indicam um nível problemático de multicolinearidade.

# ajusta o modelo para VIF  
model\_vif <- lm(PTIME ~., data = dados\_combinados)  
summary(model\_vif)

##   
## Call:  
## lm(formula = PTIME ~ ., data = dados\_combinados)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -953.13 -88.56 -5.64 99.70 701.81   
##   
## Coefficients: (4 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 153.87076 73.46943 2.094 0.036240 \*   
## DAYSWAIT\_CHRON -106.06311 6.90016 -15.371 < 2e-16 \*\*\*  
## FINAL\_MELD\_SCORE -0.77346 1.82670 -0.423 0.671993   
## AGE 66.36801 39.27170 1.690 0.091048 .   
## WGT\_KG\_TCR 0.39772 3.41706 0.116 0.907343   
## HGT\_CM\_TCR 3.34877 4.77803 0.701 0.483393   
## BMI\_TCR -0.09528 1.74655 -0.055 0.956495   
## INIT\_AGE -64.44892 39.23113 -1.643 0.100439   
## ETHCAT -2.23948 1.57455 -1.422 0.154955   
## TX\_PROCEDUR\_TY 0.11987 1.43076 0.084 0.933234   
## MED\_COND\_TRR 0.32240 1.78397 0.181 0.856589   
## AGE\_DON -0.20663 1.73014 -0.119 0.904937   
## HGT\_CM\_DON\_CALC 12.43139 5.42909 2.290 0.022044 \*   
## WGT\_KG\_DON\_CALC -28.51226 11.08105 -2.573 0.010087 \*   
## BMI\_DON\_CALC 25.08775 9.99208 2.511 0.012054 \*   
## COD\_CAD\_DON 0.58251 1.40518 0.415 0.678479   
## ETHCAT\_DON 0.05610 1.70505 0.033 0.973755   
## ABO\_MAT -3.48059 1.92780 -1.805 0.071015 .   
## COLD\_ISCH -1.67001 1.50339 -1.111 0.266654   
## HGT\_CM\_CALC -11.55649 8.22824 -1.404 0.160187   
## WGT\_KG\_CALC 22.86661 13.48813 1.695 0.090030 .   
## BMI\_CALC -17.09108 11.43594 -1.495 0.135058   
## LISTYR -290.02838 13.65404 -21.241 < 2e-16 \*\*\*  
## LiverSize NA NA NA NA   
## LiverSizeDon NA NA NA NA   
## PSTATUS0 40.87364 20.65421 1.979 0.047834 \*   
## PSTATUS1 NA NA NA NA   
## PX\_STATD -209.13442 20.88335 -10.014 < 2e-16 \*\*\*  
## ABOA1 3.03940 50.60522 0.060 0.952108   
## ABOA1B -62.21974 121.37440 -0.513 0.608218   
## ABOA2 -183.39361 120.32788 -1.524 0.127494   
## ABOA2B -416.69183 208.90473 -1.995 0.046093 \*   
## ABOAB -7.35691 15.29771 -0.481 0.630581   
## ABOB 3.75094 13.23314 0.283 0.776834   
## ABOO 5.46336 9.96950 0.548 0.583692   
## GENDER1 8.28976 4.13178 2.006 0.044832 \*   
## DIAB2 18.60190 10.87212 1.711 0.087100 .   
## DIAB3 5.09102 3.34089 1.524 0.127561   
## DIAB4 -35.58067 18.92615 -1.880 0.060125 .   
## DIAB5 -19.31849 11.35158 -1.702 0.088801 .   
## DIAB998 -15.64333 24.35366 -0.642 0.520659   
## REGION2 34.03784 24.99372 1.362 0.173257   
## REGION3 1.20278 25.18340 0.048 0.961907   
## REGION4 49.74016 30.62151 1.624 0.104315   
## REGION5 120.10284 29.40720 4.084 4.44e-05 \*\*\*  
## REGION6 111.44228 37.20997 2.995 0.002748 \*\*   
## REGION7 11.59182 28.26507 0.410 0.681728   
## REGION8 41.28242 28.78696 1.434 0.151567   
## REGION9 10.04123 25.98525 0.386 0.699189   
## REGION10 5.85453 27.06439 0.216 0.828741   
## REGION11 32.11282 25.46698 1.261 0.207337   
## PERM\_STATEAL 127.14737 53.95663 2.356 0.018458 \*   
## PERM\_STATEAR 119.95197 54.26838 2.210 0.027091 \*   
## PERM\_STATEAZ 37.11856 52.62984 0.705 0.480646   
## PERM\_STATECA 17.75909 51.88832 0.342 0.732162   
## PERM\_STATECO 102.92809 54.68392 1.882 0.059817 .   
## PERM\_STATECT 146.54982 56.80543 2.580 0.009891 \*\*   
## PERM\_STATEDC 84.18179 59.41956 1.417 0.156574   
## PERM\_STATEDE 130.07134 56.48720 2.303 0.021307 \*   
## PERM\_STATEFL 149.33238 52.88581 2.824 0.004752 \*\*   
## PERM\_STATEGA 124.98222 53.28710 2.345 0.019013 \*   
## PERM\_STATEHI 90.96553 59.02271 1.541 0.123284   
## PERM\_STATEIA 160.13565 55.63104 2.879 0.003999 \*\*   
## PERM\_STATEID 26.72636 57.25739 0.467 0.640665   
## PERM\_STATEIL 127.30176 53.68437 2.371 0.017734 \*   
## PERM\_STATEIN 151.44360 54.09070 2.800 0.005118 \*\*   
## PERM\_STATEKS 124.33624 55.22521 2.251 0.024368 \*   
## PERM\_STATEKY 155.76577 54.44917 2.861 0.004230 \*\*   
## PERM\_STATELA 185.44514 53.59875 3.460 0.000541 \*\*\*  
## PERM\_STATEMA 160.07517 56.23585 2.846 0.004424 \*\*   
## PERM\_STATEMD 140.50524 54.09852 2.597 0.009405 \*\*   
## PERM\_STATEME 125.16049 60.33298 2.074 0.038045 \*   
## PERM\_STATEMI 129.84967 54.48031 2.383 0.017161 \*   
## PERM\_STATEMN 168.32391 54.89937 3.066 0.002172 \*\*   
## PERM\_STATEMO 146.10307 54.39858 2.686 0.007241 \*\*   
## PERM\_STATEMS 178.41877 53.86704 3.312 0.000927 \*\*\*  
## PERM\_STATEMT -16.42450 64.17724 -0.256 0.798012   
## PERM\_STATENA -35.28145 56.35193 -0.626 0.531262   
## PERM\_STATENC 141.28258 54.08610 2.612 0.009003 \*\*   
## PERM\_STATEND 136.74493 65.70439 2.081 0.037426 \*   
## PERM\_STATENE 135.72787 56.59566 2.398 0.016484 \*   
## PERM\_STATENH 47.76202 60.00618 0.796 0.426069   
## PERM\_STATENJ 156.44212 54.02574 2.896 0.003787 \*\*   
## PERM\_STATENM 96.13812 55.28439 1.739 0.082053 .   
## PERM\_STATENV 19.73101 55.43116 0.356 0.721877   
## PERM\_STATENY 144.42716 54.35252 2.657 0.007884 \*\*   
## PERM\_STATEOH 149.48368 54.05844 2.765 0.005693 \*\*   
## PERM\_STATEOK 116.17387 56.45396 2.058 0.039616 \*   
## PERM\_STATEOR 65.61351 52.37203 1.253 0.210279   
## PERM\_STATEPA 126.19499 53.76590 2.347 0.018928 \*   
## PERM\_STATEPR 133.05418 55.17826 2.411 0.015902 \*   
## PERM\_STATERI 118.45175 62.85477 1.885 0.059506 .   
## PERM\_STATESC 129.97767 54.44384 2.387 0.016978 \*   
## PERM\_STATESD 71.39490 61.33026 1.164 0.244394   
## PERM\_STATETN 128.72563 53.64246 2.400 0.016417 \*   
## PERM\_STATETX 105.28627 54.61328 1.928 0.053887 .   
## PERM\_STATEUT 29.69521 55.05405 0.539 0.589628   
## PERM\_STATEVA 102.35601 53.81948 1.902 0.057205 .   
## PERM\_STATEVI 112.85203 214.42410 0.526 0.598683   
## PERM\_STATEVT 56.35049 65.40378 0.862 0.388929   
## PERM\_STATEWA 74.53313 50.62497 1.472 0.140965   
## PERM\_STATEWI 161.59474 54.72470 2.953 0.003152 \*\*   
## PERM\_STATEWV 144.47162 55.80070 2.589 0.009630 \*\*   
## PERM\_STATEWY 159.17273 67.34598 2.364 0.018111 \*   
## PERM\_STATEZZ -105.25275 116.78410 -0.901 0.367460   
## TX\_Year2002 -15.23761 217.08261 -0.070 0.944041   
## TX\_Year2003 -486.77052 103.37474 -4.709 2.51e-06 \*\*\*  
## TX\_Year2004 -413.16814 55.46821 -7.449 9.77e-14 \*\*\*  
## TX\_Year2005 -360.53181 51.65173 -6.980 3.03e-12 \*\*\*  
## TX\_Year2006 -303.34069 47.82899 -6.342 2.31e-10 \*\*\*  
## TX\_Year2007 -227.13883 44.13761 -5.146 2.68e-07 \*\*\*  
## TX\_Year2008 -122.93883 40.52849 -3.033 0.002421 \*\*   
## TX\_Year2009 -31.14601 36.76946 -0.847 0.396970   
## TX\_Year2010 48.99257 33.37940 1.468 0.142187   
## TX\_Year2011 126.54090 29.74033 4.255 2.10e-05 \*\*\*  
## TX\_Year2012 206.22889 26.60901 7.750 9.56e-15 \*\*\*  
## TX\_Year2013 317.14947 23.08687 13.737 < 2e-16 \*\*\*  
## TX\_Year2014 612.69079 18.55322 33.023 < 2e-16 \*\*\*  
## TX\_Year2015 468.13199 15.53686 30.130 < 2e-16 \*\*\*  
## TX\_Year2016 238.04021 13.25191 17.963 < 2e-16 \*\*\*  
## TX\_Year2017 32.05612 12.22905 2.621 0.008765 \*\*   
## TX\_Year2018 NA NA NA NA   
## PREV\_TXY 4.56098 7.03063 0.649 0.516519   
## GENDER\_DON1 -1.86265 3.74759 -0.497 0.619176   
## HOME\_STATE\_DONAL -8.43214 38.29102 -0.220 0.825708   
## HOME\_STATE\_DONAR 14.12669 39.35697 0.359 0.719645   
## HOME\_STATE\_DONAS -55.97579 151.54853 -0.369 0.711864   
## HOME\_STATE\_DONAZ 24.01243 36.95827 0.650 0.515882   
## HOME\_STATE\_DONCA 9.91283 35.56763 0.279 0.780475   
## HOME\_STATE\_DONCO 10.50291 39.84639 0.264 0.792102   
## HOME\_STATE\_DONCT 22.66272 42.13398 0.538 0.590670   
## HOME\_STATE\_DONDC -72.79340 48.20475 -1.510 0.131035   
## HOME\_STATE\_DONDE -57.09263 41.75440 -1.367 0.171531   
## HOME\_STATE\_DONFL 3.06359 37.00596 0.083 0.934022   
## HOME\_STATE\_DONGA 1.56225 37.58104 0.042 0.966842   
## HOME\_STATE\_DONHI 5.11009 43.00342 0.119 0.905411   
## HOME\_STATE\_DONIA -4.60050 41.75278 -0.110 0.912264   
## HOME\_STATE\_DONID -55.03713 38.34739 -1.435 0.151237   
## HOME\_STATE\_DONIL 6.17403 37.57699 0.164 0.869494   
## HOME\_STATE\_DONIN -3.61847 38.07137 -0.095 0.924280   
## HOME\_STATE\_DONKS 8.02267 39.55829 0.203 0.839288   
## HOME\_STATE\_DONKY -9.05705 38.96798 -0.232 0.816212   
## HOME\_STATE\_DONLA 4.14450 38.01969 0.109 0.913196   
## HOME\_STATE\_DONMA 12.01517 40.24625 0.299 0.765293   
## HOME\_STATE\_DONMD -17.75669 38.24346 -0.464 0.642433   
## HOME\_STATE\_DONME -1.80956 44.31471 -0.041 0.967428   
## HOME\_STATE\_DONMI 2.61737 38.32485 0.068 0.945552   
## HOME\_STATE\_DONMN 13.18706 39.53047 0.334 0.738690   
## HOME\_STATE\_DONMO 5.96244 38.08352 0.157 0.875591   
## HOME\_STATE\_DONMS 4.54109 39.33930 0.115 0.908102   
## HOME\_STATE\_DONMT 1.16415 40.88319 0.028 0.977283   
## HOME\_STATE\_DONNA -174.70614 86.42411 -2.021 0.043240 \*   
## HOME\_STATE\_DONNC -17.34720 37.97607 -0.457 0.647824   
## HOME\_STATE\_DONND 7.20380 48.67592 0.148 0.882348   
## HOME\_STATE\_DONNE -7.36642 41.80503 -0.176 0.860131   
## HOME\_STATE\_DONNH 3.39616 45.20268 0.075 0.940110   
## HOME\_STATE\_DONNJ -28.66393 38.06507 -0.753 0.451443   
## HOME\_STATE\_DONNM -0.33715 38.95924 -0.009 0.993095   
## HOME\_STATE\_DONNV 8.58283 37.46170 0.229 0.818786   
## HOME\_STATE\_DONNY -3.77070 37.74778 -0.100 0.920431   
## HOME\_STATE\_DONOH 17.74113 37.75421 0.470 0.638423   
## HOME\_STATE\_DONOK -5.47449 38.75830 -0.141 0.887676   
## HOME\_STATE\_DONOR -21.12815 35.24933 -0.599 0.548918   
## HOME\_STATE\_DONPA -16.13043 37.44103 -0.431 0.666602   
## HOME\_STATE\_DONPR -2.27250 39.45091 -0.058 0.954065   
## HOME\_STATE\_DONRI 13.13334 54.94583 0.239 0.811090   
## HOME\_STATE\_DONSC 3.87079 38.35579 0.101 0.919616   
## HOME\_STATE\_DONSD -5.66388 47.24006 -0.120 0.904567   
## HOME\_STATE\_DONTN -30.13909 37.71253 -0.799 0.424195   
## HOME\_STATE\_DONTX -4.41305 37.21778 -0.119 0.905614   
## HOME\_STATE\_DONUT 15.90617 40.17949 0.396 0.692199   
## HOME\_STATE\_DONVA -5.64629 37.94682 -0.149 0.881717   
## HOME\_STATE\_DONVI -138.15177 151.44494 -0.912 0.361661   
## HOME\_STATE\_DONVT 7.18440 54.93297 0.131 0.895947   
## HOME\_STATE\_DONWA -12.52560 31.52830 -0.397 0.691164   
## HOME\_STATE\_DONWI 3.20023 38.86388 0.082 0.934373   
## HOME\_STATE\_DONWV -5.03793 40.16520 -0.125 0.900184   
## HOME\_STATE\_DONWY 50.94196 49.81750 1.023 0.306522   
## HOME\_STATE\_DONZZ 72.53581 61.66161 1.176 0.239467   
## DIABETES\_DONU -7.61599 30.14378 -0.253 0.800537   
## DIABETES\_DONY 0.43972 4.73336 0.093 0.925986   
## HIST\_HYPERTENS\_DONU 39.99252 23.31653 1.715 0.086323 .   
## HIST\_HYPERTENS\_DONY 2.72043 3.58325 0.759 0.447737   
## ABO\_DONA1 -3.64769 4.97136 -0.734 0.463115   
## ABO\_DONA1B 12.35536 19.58785 0.631 0.528200   
## ABO\_DONA2 11.39158 9.50527 1.198 0.230755   
## ABO\_DONA2B 0.61270 23.18158 0.026 0.978914   
## ABO\_DONAB 6.39857 19.65372 0.326 0.744756   
## ABO\_DONB -2.86189 13.73652 -0.208 0.834964   
## ABO\_DONO -5.30072 10.45187 -0.507 0.612051   
## HIST\_CANCER\_DONU -14.87107 30.37012 -0.490 0.624378   
## HIST\_CANCER\_DONY -8.58887 7.93458 -1.082 0.279059   
## ALCOHOL\_HEAVY\_DONU -33.10523 11.33325 -2.921 0.003492 \*\*   
## ALCOHOL\_HEAVY\_DONY 5.85315 3.87517 1.510 0.130949   
## MALIGU -36.26514 5.68606 -6.378 1.83e-10 \*\*\*  
## MALIGY -25.44485 5.05340 -5.035 4.81e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 207.5 on 22026 degrees of freedom  
## Multiple R-squared: 0.6105, Adjusted R-squared: 0.6071   
## F-statistic: 180.7 on 191 and 22026 DF, p-value: < 2.2e-16

O modelo model\_vif apresentou uma relação linear perfeita em 4 variáveis. Temos que remover manualmente para não ocorrer o erro:Error in vif.default(model\_vif) : there are aliased coefficients in the model

# Retirando variáveis  
dados\_combinados2 <- subset(dados\_combinados, select = -c(LiverSize, LiverSizeDon, PSTATUS1, TX\_Year2018))  
model\_vif2 <- lm(PTIME ~ ., data = dados\_combinados2 )  
  
# Calcular VIF  
vif\_values <- car::vif(model\_vif2)  
str(vif\_values)

## Named num [1:191] 24.57 1.72 795.78 6.02 11.78 ...  
## - attr(\*, "names")= chr [1:191] "DAYSWAIT\_CHRON" "FINAL\_MELD\_SCORE" "AGE" "WGT\_KG\_TCR" ...

print(vif\_values)

## DAYSWAIT\_CHRON FINAL\_MELD\_SCORE AGE WGT\_KG\_TCR   
## 24.566930 1.721745 795.778986 6.024734   
## HGT\_CM\_TCR BMI\_TCR INIT\_AGE ETHCAT   
## 11.779607 1.573963 794.136026 1.279228   
## TX\_PROCEDUR\_TY MED\_COND\_TRR AGE\_DON HGT\_CM\_DON\_CALC   
## 1.056257 1.642129 1.544521 15.208550   
## WGT\_KG\_DON\_CALC BMI\_DON\_CALC COD\_CAD\_DON ETHCAT\_DON   
## 63.357062 51.516343 1.018813 1.500057   
## ABO\_MAT COLD\_ISCH HGT\_CM\_CALC WGT\_KG\_CALC   
## 1.917601 1.166214 34.933903 93.872113   
## BMI\_CALC LISTYR PSTATUS0 PX\_STATD   
## 67.480266 96.195659 43.487375 43.970257   
## ABOA1 ABOA1B ABOA2 ABOA2B   
## 1.010311 1.026277 1.008655 1.013502   
## ABOAB ABOB ABOO GENDER1   
## 5.795022 10.716269 12.656314 1.911860   
## DIAB2 DIAB3 DIAB4 DIAB5   
## 1.022689 1.096098 1.025802 1.057012   
## DIAB998 REGION2 REGION3 REGION4   
## 1.043282 34.948063 48.591641 43.845372   
## REGION5 REGION6 REGION7 REGION8   
## 50.508277 20.471720 29.160511 25.951456   
## REGION9 REGION10 REGION11 PERM\_STATEAL   
## 17.732062 32.071468 32.257286 25.123308   
## PERM\_STATEAR PERM\_STATEAZ PERM\_STATECA PERM\_STATECO   
## 16.773288 30.826398 126.444227 16.963243   
## PERM\_STATECT PERM\_STATEDC PERM\_STATEDE PERM\_STATEFL   
## 13.380340 5.151265 9.357547 91.755600   
## PERM\_STATEGA PERM\_STATEHI PERM\_STATEIA PERM\_STATEID   
## 53.178811 5.565245 12.054597 4.858900   
## PERM\_STATEIL PERM\_STATEIN PERM\_STATEKS PERM\_STATEKY   
## 60.071723 33.340459 16.677165 21.983433   
## PERM\_STATELA PERM\_STATEMA PERM\_STATEMD PERM\_STATEME   
## 27.556935 35.967223 46.207479 6.570895   
## PERM\_STATEMI PERM\_STATEMN PERM\_STATEMO PERM\_STATEMS   
## 46.021949 19.897965 31.288877 20.599344   
## PERM\_STATEMT PERM\_STATENA PERM\_STATENC PERM\_STATEND   
## 2.579559 7.049462 33.010363 2.903804   
## PERM\_STATENE PERM\_STATENH PERM\_STATENJ PERM\_STATENM   
## 9.761127 6.167690 38.033310 8.682543   
## PERM\_STATENV PERM\_STATENY PERM\_STATEOH PERM\_STATEOK   
## 7.810753 75.862256 54.917734 16.920357   
## PERM\_STATEOR PERM\_STATEPA PERM\_STATEPR PERM\_STATERI   
## 11.498622 70.798942 13.528572 4.486023   
## PERM\_STATESC PERM\_STATESD PERM\_STATETN PERM\_STATETX   
## 22.914178 3.923100 41.160568 124.532933   
## PERM\_STATEUT PERM\_STATEVA PERM\_STATEVI PERM\_STATEVT   
## 11.112308 34.418555 1.067764 3.273571   
## PERM\_STATEWA PERM\_STATEWI PERM\_STATEWV PERM\_STATEWY   
## 21.713915 29.061385 9.846035 2.525300   
## PERM\_STATEZZ TX\_Year2002 TX\_Year2003 TX\_Year2004   
## 1.266768 1.094405 1.488711 21.633742   
## TX\_Year2005 TX\_Year2006 TX\_Year2007 TX\_Year2008   
## 30.874678 27.435974 24.397436 20.896724   
## TX\_Year2009 TX\_Year2010 TX\_Year2011 TX\_Year2012   
## 18.033736 14.248404 11.037857 8.820270   
## TX\_Year2013 TX\_Year2014 TX\_Year2015 TX\_Year2016   
## 7.822466 20.246252 20.933374 17.087128   
## TX\_Year2017 PREV\_TXY GENDER\_DON1 HOME\_STATE\_DONAL   
## 9.617146 1.073308 1.732935 14.751531   
## HOME\_STATE\_DONAR HOME\_STATE\_DONAS HOME\_STATE\_DONAZ HOME\_STATE\_DONCA   
## 7.553758 1.066699 12.093489 50.048202   
## HOME\_STATE\_DONCO HOME\_STATE\_DONCT HOME\_STATE\_DONDC HOME\_STATE\_DONDE   
## 8.646025 6.549286 2.477337 4.512066   
## HOME\_STATE\_DONFL HOME\_STATE\_DONGA HOME\_STATE\_DONHI HOME\_STATE\_DONIA   
## 48.959382 25.417585 3.167645 5.272561   
## HOME\_STATE\_DONID HOME\_STATE\_DONIL HOME\_STATE\_DONIN HOME\_STATE\_DONKS   
## 2.722343 28.921013 16.805851 8.805891   
## HOME\_STATE\_DONKY HOME\_STATE\_DONLA HOME\_STATE\_DONMA HOME\_STATE\_DONMD   
## 10.642952 15.990740 12.630942 19.540734   
## HOME\_STATE\_DONME HOME\_STATE\_DONMI HOME\_STATE\_DONMN HOME\_STATE\_DONMO   
## 4.042875 23.222035 9.644368 19.473289   
## HOME\_STATE\_DONMS HOME\_STATE\_DONMT HOME\_STATE\_DONNA HOME\_STATE\_DONNC   
## 7.617482 2.052466 1.213888 21.984543   
## HOME\_STATE\_DONND HOME\_STATE\_DONNE HOME\_STATE\_DONNH HOME\_STATE\_DONNJ   
## 2.526002 5.325874 3.499927 17.089099   
## HOME\_STATE\_DONNM HOME\_STATE\_DONNV HOME\_STATE\_DONNY HOME\_STATE\_DONOH   
## 5.147581 8.152100 31.090763 28.738674   
## HOME\_STATE\_DONOK HOME\_STATE\_DONOR HOME\_STATE\_DONPA HOME\_STATE\_DONPR   
## 11.240499 4.583851 38.897325 10.451151   
## HOME\_STATE\_DONRI HOME\_STATE\_DONSC HOME\_STATE\_DONSD HOME\_STATE\_DONTN   
## 1.960775 14.144708 2.585593 22.609661   
## HOME\_STATE\_DONTX HOME\_STATE\_DONUT HOME\_STATE\_DONVA HOME\_STATE\_DONVI   
## 57.781784 5.549074 16.759870 1.065242   
## HOME\_STATE\_DONVT HOME\_STATE\_DONWA HOME\_STATE\_DONWI HOME\_STATE\_DONWV   
## 1.820031 7.236277 15.196213 6.025381   
## HOME\_STATE\_DONWY HOME\_STATE\_DONZZ DIABETES\_DONU DIABETES\_DONY   
## 1.956706 1.500010 2.602162 1.181389   
## HIST\_HYPERTENS\_DONU HIST\_HYPERTENS\_DONY ABO\_DONA1 ABO\_DONA1B   
## 1.694191 1.516828 2.035057 2.602485   
## ABO\_DONA2 ABO\_DONA2B ABO\_DONAB ABO\_DONB   
## 1.549962 1.748590 2.628748 10.523367   
## ABO\_DONO HIST\_CANCER\_DONU HIST\_CANCER\_DONY ALCOHOL\_HEAVY\_DONU   
## 14.030400 2.556642 1.052784 1.229186   
## ALCOHOL\_HEAVY\_DONY MALIGU MALIGY   
## 1.052341 4.167704 2.305228

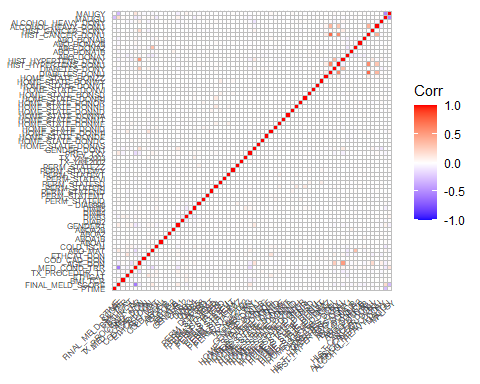
# Suponha que 'high\_vif\_names' contém os nomes das variáveis com VIF > 5  
high\_vif\_names <- names(vif\_values)[which(vif\_values > 5)]  
  
# Remover estas variáveis do dataframe 'dados\_combinados'  
dados\_combinados\_reduzido <- dados\_combinados2[, !(names(dados\_combinados2) %in% high\_vif\_names)]  
  
# Verificar a estrutura do novo dataframe para confirmar a remoção  
str(dados\_combinados\_reduzido)

## 'data.frame': 22218 obs. of 66 variables:  
## $ PTIME : num 42 739 54 127 224 547 526 315 18 717 ...  
## $ FINAL\_MELD\_SCORE : num 0.851 -0.41 -0.936 -1.041 -0.41 ...  
## $ BMI\_TCR : num -0.5983 0.0936 -0.2631 0.061 0.1383 ...  
## $ ETHCAT : num -0.533 1.638 -0.533 -0.533 -0.533 ...  
## $ TX\_PROCEDUR\_TY : num -0.113 -0.113 -0.113 -0.113 -0.113 ...  
## $ MED\_COND\_TRR : num 0.574 0.574 0.574 0.574 0.574 ...  
## $ AGE\_DON : num -0.826 -1.254 -1.132 0.151 1.679 ...  
## $ COD\_CAD\_DON : num 6.87 -0.138 -0.138 -0.145 -0.145 ...  
## $ ETHCAT\_DON : num -0.546 -0.546 -0.546 -0.546 -0.546 ...  
## $ ABO\_MAT : num -0.244 -0.244 -0.244 -0.244 -0.244 ...  
## $ COLD\_ISCH : num 1.42 1.42 1.04 -1.21 1.15 ...  
## $ ABOA1 : num 0 0 0 0 0 0 1 0 0 0 ...  
## $ ABOA1B : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ ABOA2 : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ ABOA2B : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ GENDER1 : num 1 0 0 0 0 0 0 0 1 0 ...  
## $ DIAB2 : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DIAB3 : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DIAB4 : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DIAB5 : num 0 0 0 0 1 0 0 1 0 0 ...  
## $ DIAB998 : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ PERM\_STATEID : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ PERM\_STATEMT : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ PERM\_STATEND : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ PERM\_STATERI : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ PERM\_STATESD : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ PERM\_STATEVI : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ PERM\_STATEVT : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ PERM\_STATEWY : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ PERM\_STATEZZ : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ TX\_Year2002 : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ TX\_Year2003 : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ PREV\_TXY : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ GENDER\_DON1 : num 0 0 1 1 0 1 0 0 0 1 ...  
## $ HOME\_STATE\_DONAS : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ HOME\_STATE\_DONDC : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ HOME\_STATE\_DONDE : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ HOME\_STATE\_DONHI : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ HOME\_STATE\_DONID : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ HOME\_STATE\_DONME : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ HOME\_STATE\_DONMT : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ HOME\_STATE\_DONNA : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ HOME\_STATE\_DONND : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ HOME\_STATE\_DONNH : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ HOME\_STATE\_DONOR : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ HOME\_STATE\_DONRI : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ HOME\_STATE\_DONSD : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ HOME\_STATE\_DONVI : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ HOME\_STATE\_DONVT : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ HOME\_STATE\_DONWY : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ HOME\_STATE\_DONZZ : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DIABETES\_DONU : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DIABETES\_DONY : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ HIST\_HYPERTENS\_DONU: num 0 0 0 0 0 0 0 0 0 0 ...  
## $ HIST\_HYPERTENS\_DONY: num 0 0 0 1 0 0 1 1 0 0 ...  
## $ ABO\_DONA1 : num 0 0 1 0 0 0 0 0 0 0 ...  
## $ ABO\_DONA1B : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ ABO\_DONA2 : num 0 0 0 0 1 0 1 0 0 0 ...  
## $ ABO\_DONA2B : num 0 0 0 0 0 0 0 0 1 0 ...  
## $ ABO\_DONAB : num 0 1 0 1 0 0 0 0 0 0 ...  
## $ HIST\_CANCER\_DONU : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ HIST\_CANCER\_DONY : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ ALCOHOL\_HEAVY\_DONU : num 0 0 0 0 1 0 0 0 0 0 ...  
## $ ALCOHOL\_HEAVY\_DONY : num 0 0 0 1 0 0 0 1 0 0 ...  
## $ MALIGU : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ MALIGY : num 1 0 0 0 0 1 0 0 0 0 ...

View(dados\_combinados\_reduzido)

**Correlação entre as variáveis**

df\_corr <- round(cor(dados\_combinados\_reduzido, use = "complete.obs"), 2)  
  
corr\_plot <- ggcorrplot(df\_corr, tl.cex = 6)   
  
print(corr\_plot)



ggsave("Health\_analytics/corr\_plot.png", corr\_plot, width = 10, height = 8)

#### Modelagem Preditiva com Modelo de Regressão Linear

Vamos trabalhar apenas com as variáveis mais significativas para o problema por meio da interpretação do p-valor.

modelo\_lm <- lm(PTIME ~ .,data = dados\_combinados\_reduzido)  
  
# Obter o resumo do modelo  
summary(modelo\_lm)

##   
## Call:  
## lm(formula = PTIME ~ ., data = dados\_combinados\_reduzido)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -750.43 -262.78 -24.76 288.73 764.47   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 650.7058 5.4184 120.093 < 2e-16 \*\*\*  
## FINAL\_MELD\_SCORE -4.8161 2.6789 -1.798 0.072226 .   
## BMI\_TCR -0.5927 2.1126 -0.281 0.779042   
## ETHCAT -2.5160 2.1405 -1.175 0.239845   
## TX\_PROCEDUR\_TY -0.6256 2.1342 -0.293 0.769426   
## MED\_COND\_TRR -10.5474 2.6165 -4.031 5.57e-05 \*\*\*  
## AGE\_DON -7.0885 2.5548 -2.775 0.005532 \*\*   
## COD\_CAD\_DON -0.5792 2.1046 -0.275 0.783140   
## ETHCAT\_DON 0.5058 2.1711 0.233 0.815800   
## ABO\_MAT -3.6493 2.2526 -1.620 0.105249   
## COLD\_ISCH -4.0884 2.1351 -1.915 0.055523 .   
## ABOA1 12.7510 75.9597 0.168 0.866691   
## ABOA1B -142.6127 180.8977 -0.788 0.430494   
## ABOA2 -63.6890 180.7328 -0.352 0.724547   
## ABOA2B -563.7049 313.4957 -1.798 0.072171 .   
## GENDER1 0.9168 4.5978 0.199 0.841944   
## DIAB2 -17.3274 16.2940 -1.063 0.287603   
## DIAB3 2.4066 4.8817 0.493 0.622036   
## DIAB4 -96.1653 28.2429 -3.405 0.000663 \*\*\*  
## DIAB5 -49.5260 16.8430 -2.940 0.003281 \*\*   
## DIAB998 21.0231 36.0892 0.583 0.560214   
## PERM\_STATEID 35.2496 39.2884 0.897 0.369621   
## PERM\_STATEMT -33.8320 60.3274 -0.561 0.574935   
## PERM\_STATEND 11.3342 58.2974 0.194 0.845848   
## PERM\_STATERI -46.7993 45.4450 -1.030 0.303115   
## PERM\_STATESD -25.0407 47.0166 -0.533 0.594321   
## PERM\_STATEVI -79.3002 313.0158 -0.253 0.800006   
## PERM\_STATEVT -45.4969 54.7377 -0.831 0.405881   
## PERM\_STATEWY 47.8162 64.0029 0.747 0.455014   
## PERM\_STATEZZ -190.0967 156.7742 -1.213 0.225315   
## TX\_Year2002 541.7204 313.2216 1.730 0.083731 .   
## TX\_Year2003 -149.9915 128.3352 -1.169 0.242518   
## PREV\_TXY -22.0148 10.4529 -2.106 0.035207 \*   
## GENDER\_DON1 1.5115 4.4336 0.341 0.733159   
## HOME\_STATE\_DONAS 26.9730 221.3170 0.122 0.902999   
## HOME\_STATE\_DONDC -166.1459 46.2654 -3.591 0.000330 \*\*\*  
## HOME\_STATE\_DONDE -55.2830 29.6792 -1.863 0.062520 .   
## HOME\_STATE\_DONHI 76.6798 36.9608 2.075 0.038033 \*   
## HOME\_STATE\_DONID -40.3138 35.1662 -1.146 0.251650   
## HOME\_STATE\_DONME -9.5080 33.4359 -0.284 0.776134   
## HOME\_STATE\_DONMT 5.3205 43.0878 0.123 0.901728   
## HOME\_STATE\_DONNA -213.3309 118.3498 -1.803 0.071473 .   
## HOME\_STATE\_DONND -21.2975 46.3500 -0.459 0.645885   
## HOME\_STATE\_DONNH 9.7733 36.7393 0.266 0.790228   
## HOME\_STATE\_DONOR 37.9365 24.9138 1.523 0.127845   
## HOME\_STATE\_DONRI 34.0401 59.5548 0.572 0.567615   
## HOME\_STATE\_DONSD -69.6256 44.4808 -1.565 0.117528   
## HOME\_STATE\_DONVI -80.9244 221.3743 -0.366 0.714701   
## HOME\_STATE\_DONVT -44.4469 61.4672 -0.723 0.469626   
## HOME\_STATE\_DONWY 121.1846 53.9578 2.246 0.024719 \*   
## HOME\_STATE\_DONZZ 19.4810 76.3359 0.255 0.798570   
## DIABETES\_DONU 44.9839 45.3043 0.993 0.320757   
## DIABETES\_DONY 3.4540 7.0832 0.488 0.625820   
## HIST\_HYPERTENS\_DONU 57.0981 35.0436 1.629 0.103254   
## HIST\_HYPERTENS\_DONY 2.4073 5.3007 0.454 0.649730   
## ABO\_DONA1 -4.4075 5.3546 -0.823 0.410441   
## ABO\_DONA1B 26.7270 18.4248 1.451 0.146905   
## ABO\_DONA2 1.9067 12.2343 0.156 0.876153   
## ABO\_DONA2B 11.6966 26.5234 0.441 0.659224   
## ABO\_DONAB -13.5193 18.3997 -0.735 0.462496   
## HIST\_CANCER\_DONU -32.4652 45.6564 -0.711 0.477045   
## HIST\_CANCER\_DONY -9.3658 11.9141 -0.786 0.431810   
## ALCOHOL\_HEAVY\_DONU -50.4120 17.0160 -2.963 0.003053 \*\*   
## ALCOHOL\_HEAVY\_DONY 6.7008 5.8059 1.154 0.248456   
## MALIGU -254.2495 5.0176 -50.672 < 2e-16 \*\*\*  
## MALIGY -146.9076 6.0941 -24.106 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 312.9 on 22152 degrees of freedom  
## Multiple R-squared: 0.1092, Adjusted R-squared: 0.1065   
## F-statistic: 41.76 on 65 and 22152 DF, p-value: < 2.2e-16

**Extração das varíáveis mais significativas do modelo**

# Extrair p-valores  
p\_values <- coef(summary(modelo\_lm))[, "Pr(>|t|)"]  
  
# Filtrar variáveis com p-valor < 0.05  
var\_significativas <- names(p\_values)[p\_values < 0.05]  
var\_significativas <- setdiff(var\_significativas, "(Intercept)")  
var\_significativas <- c("PTIME", var\_significativas)  
print(var\_significativas)

## [1] "PTIME" "MED\_COND\_TRR" "AGE\_DON"   
## [4] "DIAB4" "DIAB5" "PREV\_TXY"   
## [7] "HOME\_STATE\_DONDC" "HOME\_STATE\_DONHI" "HOME\_STATE\_DONWY"   
## [10] "ALCOHOL\_HEAVY\_DONU" "MALIGU" "MALIGY"

**Divisão dos dados em treino e teste**

dados\_final <- dados\_combinados\_reduzido[, var\_significativas]  
View(dados\_final)  
  
# Divisão dos dados em treino e teste  
set.seed(1)  
index <- sample(1:nrow(dados\_final), dim(dados\_final)[1]\*.7)  
dados\_treino <- dados\_final[index,]  
dados\_teste <- dados\_final[-index,]  
  
View(dados\_treino )  
View(dados\_teste)

* 1. **- Treinando o primeiro modelo**

# Instalando e carregando o pacote necessário para RMSE e MAE  
if (!require(Metrics)) {  
 install.packages("Metrics")  
 library(Metrics)  
}

## Carregando pacotes exigidos: Metrics

## Warning: package 'Metrics' was built under R version 4.3.3

##   
## Attaching package: 'Metrics'

## The following objects are masked from 'package:caret':  
##   
## precision, recall

## The following object is masked from 'package:forecast':  
##   
## accuracy

# Treinando o modelo  
modelo\_v1 <- lm(PTIME ~ ., data = dados\_treino)  
  
# Fazendo previsões com dados de treino  
modelo\_v1\_pred\_1 <- predict(modelo\_v1, newdata = dados\_treino)  
  
# Calculando RMSE para dados de treino  
rmse\_treino <- rmse(dados\_treino$PTIME, modelo\_v1\_pred\_1)  
print(paste("RMSE Treino:", rmse\_treino))

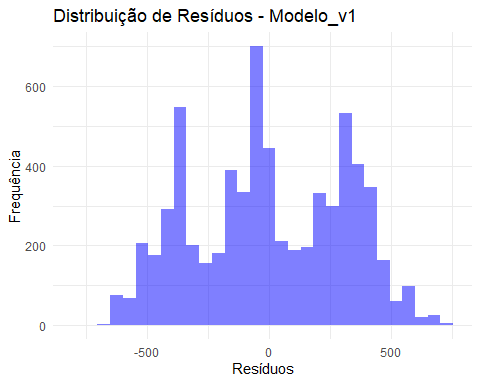
## [1] "RMSE Treino: 314.456171723449"

# Fazendo previsões com dados de teste  
modelo\_v1\_pred\_2 <- predict(modelo\_v1, newdata = dados\_teste)  
  
# Calculando RMSE para dados de teste  
rmse\_teste <- rmse(dados\_teste$PTIME, modelo\_v1\_pred\_2)  
print(paste("RMSE Teste:", rmse\_teste))

## [1] "RMSE Teste: 309.120918013813"

Plot dos Resíduos

# Adicionando os resíduos ao dataframe  
dados\_teste$residuos <- dados\_teste$PTIME - modelo\_v1\_pred\_2  
  
# Carregando o pacote ggplot2  
library(ggplot2)  
  
# Criando o histograma de resíduos  
plot\_gg <- ggplot(dados\_teste, aes(x=residuos)) +  
 geom\_histogram(fill="blue", alpha=0.5, bins=30) + # Definindo o número de bins  
 labs(title="Distribuição de Resíduos - Modelo\_v1", x="Resíduos", y="Frequência") +  
 theme\_minimal()  
  
# Exibe o gráfico no RStudio  
print(plot\_gg)



# Salva o gráfico no arquivo usando ggsave  
ggsave("Distribuição de Resíduos Modelo\_v1.png", plot = plot\_gg, width=10, height=8, units="cm")

* Centralização: Os resíduos parecem estar centrados em torno de zero, o que é um bom sinal. Um modelo bem ajustado terá seus resíduos distribuídos de forma simétrica ao redor do zero.
* Forma da Distribuição: A distribuição dos resíduos não parece ser perfeitamente normal. Há várias modas visíveis (picos na distribuição), sugerindo que a distribuição dos resíduos é multimodal.
* Variância dos Resíduos: Há uma variação considerável nos resíduos, com alguns deles estendendo-se para além de -500 e 500. Isso pode indicar que o modelo tem dificuldade em capturar todos os padrões nos dados, resultando em previsões que, em alguns casos, estão bastante distantes dos valores reais.
* Homoscedasticidade: Idealmente, queremos que os resíduos sejam homoscedásticos, o que significa que os resíduos têm variâncias constantes em toda a gama de previsões. No entanto, sem uma visualização de um gráfico de dispersão dos resíduos contra os valores previstos, não podemos avaliar completamente a homoscedasticidade. Se os resíduos mostrarem padrões claros ou se expandirem (em forma de funil, por exemplo) à medida que os valores ajustados aumentam ou diminuem, isso indicaria heteroscedasticidade.
* Outliers: A presença de resíduos com valores altos (tanto positivos quanto negativos) pode indicar outliers no dataset ou que o modelo não está lidando bem com todas as nuances dos dados.

**1.7 - Treinando o segundo modelo**

Usando regularização LASSO para seleção de variáveis

# Separando a variável alvo das preditoras  
X <- as.matrix(subset(dados\_combinados, select = -PTIME))  
  
# Garanta que 'y' não seja padronizado junto com as variáveis preditoras  
Y <- dados\_num$PTIME   
  
# Para reprodutibilidade  
set.seed(123)   
  
# Ajuste do modelo LASSO usando validação cruzada. Alpha = 1 indica LASSO  
cv\_model <- cv.glmnet(X, Y, alpha = 1)

**Nota**: Sobre o uso de glmnet, o objeto x usado em glmnet deve ser uma matriz e y um vetor. Trabalhando com um dataframe em R, deve-se converter apropriadamente o dataframe para matriz antes de aplicar glmnet.

Extraindo os coeficientes e verificando as variáveis selecionadas

# Verifique se o cv\_model possui os componentes esperados  
print(cv\_model$lambda.min) # Deve retornar o valor de lambda que minimiza o MSE

## [1] 0.1277909

print(cv\_model$lambda.1se) # Deve retornar o valor de lambda dentro de um SE do mínimo

## [1] 1.307979

# Extraia os coeficientes para lambda.min e lambda.1se  
coeficientes\_min <- coef(cv\_model, s = "lambda.min")  
coeficientes\_1se <- coef(cv\_model, s = "lambda.1se")  
  
# Verifique se os coeficientes são NULL  
print(coeficientes\_min)

## 196 x 1 sparse Matrix of class "dgCMatrix"  
## s1  
## (Intercept) 494.7634771  
## DAYSWAIT\_CHRON -78.9300396  
## FINAL\_MELD\_SCORE -0.3741940  
## AGE 1.7540775  
## WGT\_KG\_TCR 0.3012570  
## HGT\_CM\_TCR 1.7461884  
## BMI\_TCR -0.2202347  
## INIT\_AGE .   
## ETHCAT -1.9800508  
## TX\_PROCEDUR\_TY 0.1380075  
## MED\_COND\_TRR 0.1223583  
## AGE\_DON .   
## HGT\_CM\_DON\_CALC 1.2830132  
## WGT\_KG\_DON\_CALC -4.6107975  
## BMI\_DON\_CALC 3.8622525  
## COD\_CAD\_DON 0.4754528  
## ETHCAT\_DON .   
## ABO\_MAT -3.6276031  
## COLD\_ISCH -1.6831946  
## HGT\_CM\_CALC .   
## WGT\_KG\_CALC 3.2341653  
## BMI\_CALC .   
## LISTYR -235.4143684  
## LiverSize .   
## LiverSizeDon -0.2713478  
## PSTATUS0 41.0900561  
## PSTATUS1 .   
## PX\_STATD -208.7487719  
## ABOA1 .   
## ABOA1B -54.5823902  
## ABOA2 -164.0083311  
## ABOA2B -395.1001177  
## ABOAB -6.1214401  
## ABOB 0.8335776  
## ABOO 0.9272885  
## GENDER1 7.3633253  
## DIAB2 17.0279056  
## DIAB3 4.8370996  
## DIAB4 -35.1683014  
## DIAB5 -18.7534188  
## DIAB998 -14.0959244  
## REGION2 4.8779486  
## REGION3 -24.2470805  
## REGION4 6.9134816  
## REGION5 77.9241215  
## REGION6 45.6902544  
## REGION7 -8.9181444  
## REGION8 15.3655429  
## REGION9 -8.0256665  
## REGION10 -14.1509426  
## REGION11 3.8944009  
## PERM\_STATEAL .   
## PERM\_STATEAR -2.7866500  
## PERM\_STATEAZ -71.7412848  
## PERM\_STATECA -93.2101088  
## PERM\_STATECO -18.0725027  
## PERM\_STATECT 5.0848295  
## PERM\_STATEDC -40.6167361  
## PERM\_STATEDE .   
## PERM\_STATEFL 22.9512912  
## PERM\_STATEGA .   
## PERM\_STATEHI -2.7999837  
## PERM\_STATEIA 32.5816808  
## PERM\_STATEID -78.0275023  
## PERM\_STATEIL .   
## PERM\_STATEIN 20.7420703  
## PERM\_STATEKS .   
## PERM\_STATEKY 27.8630837  
## PERM\_STATELA 60.8581864  
## PERM\_STATEMA 15.0916120  
## PERM\_STATEMD 10.7697205  
## PERM\_STATEME -16.2904763  
## PERM\_STATEMI .   
## PERM\_STATEMN 39.5401668  
## PERM\_STATEMO 21.9213219  
## PERM\_STATEMS 52.5853248  
## PERM\_STATEMT -121.3751135  
## PERM\_STATENA -160.4458913  
## PERM\_STATENC 12.9992049  
## PERM\_STATEND 2.6649305  
## PERM\_STATENE 9.7580618  
## PERM\_STATENH -92.0689451  
## PERM\_STATENJ 25.4016052  
## PERM\_STATENM -18.3053615  
## PERM\_STATENV -88.3308966  
## PERM\_STATENY 9.3450845  
## PERM\_STATEOH 18.0975849  
## PERM\_STATEOK 2.4238714  
## PERM\_STATEOR -29.3087536  
## PERM\_STATEPA -1.3848604  
## PERM\_STATEPR 4.9921903  
## PERM\_STATERI -21.8672637  
## PERM\_STATESC 3.6736163  
## PERM\_STATESD -50.8871795  
## PERM\_STATETN 1.8277160  
## PERM\_STATETX -7.6223775  
## PERM\_STATEUT -78.6644208  
## PERM\_STATEVA -24.9488566  
## PERM\_STATEVI .   
## PERM\_STATEVT -80.1322473  
## PERM\_STATEWA -15.4265506  
## PERM\_STATEWI 31.9575273  
## PERM\_STATEWV 13.5995321  
## PERM\_STATEWY 36.6409853  
## PERM\_STATEZZ -211.3694928  
## TX\_Year2002 .   
## TX\_Year2003 -499.0197195  
## TX\_Year2004 -445.5162589  
## TX\_Year2005 -406.7754619  
## TX\_Year2006 -366.2683213  
## TX\_Year2007 -305.3069110  
## TX\_Year2008 -216.4401061  
## TX\_Year2009 -141.4727448  
## TX\_Year2010 -76.9613406  
## TX\_Year2011 -16.1176115  
## TX\_Year2012 47.5631688  
## TX\_Year2013 141.7945893  
## TX\_Year2014 421.6061642  
## TX\_Year2015 262.2597343  
## TX\_Year2016 16.0850187  
## TX\_Year2017 -203.6870642  
## TX\_Year2018 -245.0788541  
## PREV\_TXY 3.0675867  
## GENDER\_DON1 -0.8502094  
## HOME\_STATE\_DONAL -7.9017024  
## HOME\_STATE\_DONAR 11.5610355  
## HOME\_STATE\_DONAS -43.4757244  
## HOME\_STATE\_DONAZ 22.4585728  
## HOME\_STATE\_DONCA 9.6429215  
## HOME\_STATE\_DONCO 4.2837732  
## HOME\_STATE\_DONCT 10.3097450  
## HOME\_STATE\_DONDC -65.3097884  
## HOME\_STATE\_DONDE -49.8792178  
## HOME\_STATE\_DONFL 1.2488467  
## HOME\_STATE\_DONGA .   
## HOME\_STATE\_DONHI 9.0791758  
## HOME\_STATE\_DONIA -5.0167308  
## HOME\_STATE\_DONID -47.4831852  
## HOME\_STATE\_DONIL 1.6321119  
## HOME\_STATE\_DONIN -5.5303102  
## HOME\_STATE\_DONKS 4.5504489  
## HOME\_STATE\_DONKY -5.2225761  
## HOME\_STATE\_DONLA 0.8954527  
## HOME\_STATE\_DONMA 1.7391023  
## HOME\_STATE\_DONMD -11.4369389  
## HOME\_STATE\_DONME -8.4365213  
## HOME\_STATE\_DONMI .   
## HOME\_STATE\_DONMN 7.7702676  
## HOME\_STATE\_DONMO 4.0580174  
## HOME\_STATE\_DONMS 3.3408797  
## HOME\_STATE\_DONMT .   
## HOME\_STATE\_DONNA -174.4401309  
## HOME\_STATE\_DONNC -13.1465836  
## HOME\_STATE\_DONND .   
## HOME\_STATE\_DONNE -8.2439147  
## HOME\_STATE\_DONNH -4.2072745  
## HOME\_STATE\_DONNJ -22.9517094  
## HOME\_STATE\_DONNM .   
## HOME\_STATE\_DONNV 7.5282746  
## HOME\_STATE\_DONNY -3.3504404  
## HOME\_STATE\_DONOH 16.2066563  
## HOME\_STATE\_DONOK -1.3856274  
## HOME\_STATE\_DONOR -8.6090949  
## HOME\_STATE\_DONPA -11.1167900  
## HOME\_STATE\_DONPR -0.6553117  
## HOME\_STATE\_DONRI 2.9843438  
## HOME\_STATE\_DONSC 5.5719612  
## HOME\_STATE\_DONSD -6.6749427  
## HOME\_STATE\_DONTN -26.8236666  
## HOME\_STATE\_DONTX -1.1731629  
## HOME\_STATE\_DONUT 13.5319461  
## HOME\_STATE\_DONVA -1.4119102  
## HOME\_STATE\_DONVI -124.6119571  
## HOME\_STATE\_DONVT 0.0624090  
## HOME\_STATE\_DONWA -8.4554974  
## HOME\_STATE\_DONWI .   
## HOME\_STATE\_DONWV .   
## HOME\_STATE\_DONWY 45.5740538  
## HOME\_STATE\_DONZZ 67.4755851  
## DIABETES\_DONU -4.1661385  
## DIABETES\_DONY 0.1662661  
## HIST\_HYPERTENS\_DONU 35.0478986  
## HIST\_HYPERTENS\_DONY 1.7807818  
## ABO\_DONA1 -3.1950724  
## ABO\_DONA1B 10.7488960  
## ABO\_DONA2 12.0773624  
## ABO\_DONA2B .   
## ABO\_DONAB 3.7127820  
## ABO\_DONB .   
## ABO\_DONO -0.5034921  
## HIST\_CANCER\_DONU -13.4322478  
## HIST\_CANCER\_DONY -8.3333125  
## ALCOHOL\_HEAVY\_DONU -31.6181662  
## ALCOHOL\_HEAVY\_DONY 5.6160619  
## MALIGU -35.7083510  
## MALIGY -24.4486939

print(coeficientes\_1se)

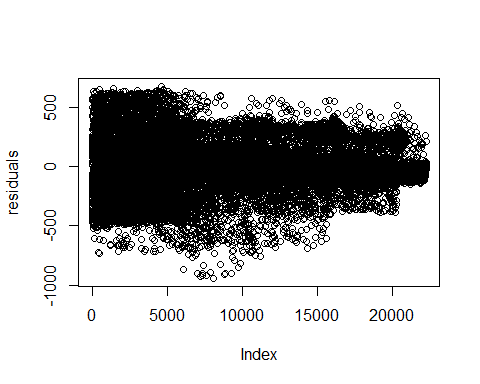
## 196 x 1 sparse Matrix of class "dgCMatrix"  
## s1  
## (Intercept) 5.599636e+02  
## DAYSWAIT\_CHRON -1.899756e+01  
## FINAL\_MELD\_SCORE .   
## AGE .   
## WGT\_KG\_TCR .   
## HGT\_CM\_TCR .   
## BMI\_TCR .   
## INIT\_AGE 3.333456e-01  
## ETHCAT -9.556488e-01  
## TX\_PROCEDUR\_TY .   
## MED\_COND\_TRR .   
## AGE\_DON .   
## HGT\_CM\_DON\_CALC .   
## WGT\_KG\_DON\_CALC .   
## BMI\_DON\_CALC .   
## COD\_CAD\_DON .   
## ETHCAT\_DON .   
## ABO\_MAT -2.255133e+00  
## COLD\_ISCH -5.446566e-01  
## HGT\_CM\_CALC .   
## WGT\_KG\_CALC 2.370060e+00  
## BMI\_CALC .   
## LISTYR -6.939585e+01  
## LiverSize .   
## LiverSizeDon .   
## PSTATUS0 3.818183e+01  
## PSTATUS1 -7.607527e-09  
## PX\_STATD -1.953757e+02  
## ABOA1 .   
## ABOA1B .   
## ABOA2 -6.175340e+01  
## ABOA2B -2.020890e+02  
## ABOAB .   
## ABOB .   
## ABOO .   
## GENDER1 1.175920e+00  
## DIAB2 5.950624e+00  
## DIAB3 2.568680e+00  
## DIAB4 -2.408634e+01  
## DIAB5 -9.603548e+00  
## DIAB998 .   
## REGION2 .   
## REGION3 -1.703917e+01  
## REGION4 .   
## REGION5 1.898953e+01  
## REGION6 5.231666e+00  
## REGION7 .   
## REGION8 1.165590e+01  
## REGION9 .   
## REGION10 .   
## REGION11 .   
## PERM\_STATEAL .   
## PERM\_STATEAR .   
## PERM\_STATEAZ .   
## PERM\_STATECA -3.253641e+01  
## PERM\_STATECO -2.629614e+00  
## PERM\_STATECT .   
## PERM\_STATEDC -2.015197e+01  
## PERM\_STATEDE .   
## PERM\_STATEFL 1.159854e+01  
## PERM\_STATEGA .   
## PERM\_STATEHI 1.350374e+01  
## PERM\_STATEIA 1.874800e+01  
## PERM\_STATEID .   
## PERM\_STATEIL .   
## PERM\_STATEIN .   
## PERM\_STATEKS .   
## PERM\_STATEKY 1.637023e+01  
## PERM\_STATELA 4.845801e+01  
## PERM\_STATEMA 6.227478e+00  
## PERM\_STATEMD .   
## PERM\_STATEME .   
## PERM\_STATEMI -6.373126e+00  
## PERM\_STATEMN 2.154607e+01  
## PERM\_STATEMO 2.257922e+01  
## PERM\_STATEMS 3.982479e+01  
## PERM\_STATEMT -5.544062e+01  
## PERM\_STATENA -1.437162e+02  
## PERM\_STATENC .   
## PERM\_STATEND .   
## PERM\_STATENE .   
## PERM\_STATENH -7.318732e+01  
## PERM\_STATENJ 7.809606e+00  
## PERM\_STATENM .   
## PERM\_STATENV -7.900050e+00  
## PERM\_STATENY .   
## PERM\_STATEOH 1.309749e+00  
## PERM\_STATEOK .   
## PERM\_STATEOR .   
## PERM\_STATEPA -3.175072e+00  
## PERM\_STATEPR .   
## PERM\_STATERI .   
## PERM\_STATESC .   
## PERM\_STATESD -2.416347e+01  
## PERM\_STATETN .   
## PERM\_STATETX .   
## PERM\_STATEUT -6.869507e-02  
## PERM\_STATEVA -2.024483e+01  
## PERM\_STATEVI .   
## PERM\_STATEVT -4.886041e+01  
## PERM\_STATEWA 7.670614e+00  
## PERM\_STATEWI 1.384013e+01  
## PERM\_STATEWV .   
## PERM\_STATEWY 2.479133e+01  
## PERM\_STATEZZ -1.185527e+02  
## TX\_Year2002 2.810411e+02  
## TX\_Year2003 .   
## TX\_Year2004 -4.151918e+01  
## TX\_Year2005 -4.508547e+01  
## TX\_Year2006 -5.392957e+01  
## TX\_Year2007 -4.005926e+01  
## TX\_Year2008 .   
## TX\_Year2009 1.123117e+01  
## TX\_Year2010 2.891679e+01  
## TX\_Year2011 3.775633e+01  
## TX\_Year2012 5.965281e+01  
## TX\_Year2013 1.078857e+02  
## TX\_Year2014 3.508550e+02  
## TX\_Year2015 1.487811e+02  
## TX\_Year2016 -1.397861e+02  
## TX\_Year2017 -4.032219e+02  
## TX\_Year2018 -4.674448e+02  
## PREV\_TXY .   
## GENDER\_DON1 .   
## HOME\_STATE\_DONAL -3.044150e+00  
## HOME\_STATE\_DONAR .   
## HOME\_STATE\_DONAS .   
## HOME\_STATE\_DONAZ 8.367062e+00  
## HOME\_STATE\_DONCA 1.951906e+00  
## HOME\_STATE\_DONCO .   
## HOME\_STATE\_DONCT .   
## HOME\_STATE\_DONDC -3.347179e+01  
## HOME\_STATE\_DONDE -2.209955e+01  
## HOME\_STATE\_DONFL .   
## HOME\_STATE\_DONGA .   
## HOME\_STATE\_DONHI 7.121927e+00  
## HOME\_STATE\_DONIA .   
## HOME\_STATE\_DONID -2.781549e+01  
## HOME\_STATE\_DONIL .   
## HOME\_STATE\_DONIN .   
## HOME\_STATE\_DONKS .   
## HOME\_STATE\_DONKY .   
## HOME\_STATE\_DONLA .   
## HOME\_STATE\_DONMA .   
## HOME\_STATE\_DONMD .   
## HOME\_STATE\_DONME .   
## HOME\_STATE\_DONMI .   
## HOME\_STATE\_DONMN .   
## HOME\_STATE\_DONMO 1.029925e+00  
## HOME\_STATE\_DONMS .   
## HOME\_STATE\_DONMT .   
## HOME\_STATE\_DONNA -1.078445e+02  
## HOME\_STATE\_DONNC .   
## HOME\_STATE\_DONND .   
## HOME\_STATE\_DONNE .   
## HOME\_STATE\_DONNH .   
## HOME\_STATE\_DONNJ -4.183570e+00  
## HOME\_STATE\_DONNM .   
## HOME\_STATE\_DONNV .   
## HOME\_STATE\_DONNY .   
## HOME\_STATE\_DONOH 1.039290e+01  
## HOME\_STATE\_DONOK .   
## HOME\_STATE\_DONOR .   
## HOME\_STATE\_DONPA .   
## HOME\_STATE\_DONPR .   
## HOME\_STATE\_DONRI .   
## HOME\_STATE\_DONSC 3.931711e+00  
## HOME\_STATE\_DONSD .   
## HOME\_STATE\_DONTN -1.219628e+01  
## HOME\_STATE\_DONTX .   
## HOME\_STATE\_DONUT .   
## HOME\_STATE\_DONVA .   
## HOME\_STATE\_DONVI .   
## HOME\_STATE\_DONVT .   
## HOME\_STATE\_DONWA .   
## HOME\_STATE\_DONWI .   
## HOME\_STATE\_DONWV .   
## HOME\_STATE\_DONWY 1.591299e+01  
## HOME\_STATE\_DONZZ 2.632259e+01  
## DIABETES\_DONU .   
## DIABETES\_DONY .   
## HIST\_HYPERTENS\_DONU 6.329774e+00  
## HIST\_HYPERTENS\_DONY .   
## ABO\_DONA1 -1.092162e-01  
## ABO\_DONA1B .   
## ABO\_DONA2 2.863740e+00  
## ABO\_DONA2B .   
## ABO\_DONAB .   
## ABO\_DONB .   
## ABO\_DONO .   
## HIST\_CANCER\_DONU .   
## HIST\_CANCER\_DONY .   
## ALCOHOL\_HEAVY\_DONU -2.170654e+01  
## ALCOHOL\_HEAVY\_DONY 2.790291e+00  
## MALIGU -3.618465e+01  
## MALIGY -2.042191e+01

Utilizaremos lambda\_1se para treinar o modelo, que é o valor de lambda mais simples (maior) que está dentro de um desvio padrão do erro mínimo. Este lambda geralmente leva a modelos mais simples e potencialmente mais robustos.

lambda\_1se <- cv\_model$lambda.1se  
  
# Usar esse lambda para treinar o modelo  
model1 <- glmnet(X, Y, alpha = 1, lambda = lambda\_1se)

Previsões do modelo

predicted\_values <- predict(model1, X, s = lambda\_1se)  
residuals <- Y - predicted\_values  
plot(residuals)



**Medidas de Ajuste e Performance**

rmse\_model1 <- rmse(Y, predicted\_values)  
print(paste("RMSE model1:", rmse\_model1))

## [1] "RMSE model1: 208.421051823622"

**1.8 - Treinando o terceiro modelo**

# Carregar a biblioteca necessária  
library(caret)  
  
# Para reprodutibilidade  
set.seed(123)   
  
# Criar índices para a divisão  
trainIndex <- createDataPartition(dados\_combinados$PTIME, p = 0.8, list = FALSE)  
  
# Dividir os dados em treino e teste  
trainData <- dados\_combinados[trainIndex, ]  
testData <- dados\_combinados[-trainIndex, ]

Ajustar o modelo LASSO no conjunto de treino

x\_train <- model.matrix(~ . -1, data = trainData[, -which(names(trainData) == "PTIME")])  
y\_train <- trainData$PTIME  
  
cv\_model <- cv.glmnet(x\_train, y\_train, alpha = 1)  
  
# Avaliar o modelo no conjunto de teste  
x\_test <- model.matrix(~ . -1, data = testData[, -which(names(testData) == "PTIME")])  
y\_test <- testData$PTIME  
  
predictions <- predict(cv\_model, s = "lambda.1se", newx = x\_test)  
rmse\_test <- sqrt(mean((predictions - y\_test)^2))  
print(rmse\_test)

## [1] 208.848

**1.9 - Treinando o quarto modelo**

**Dividindo em em dados de treino**

#padronizando a variável alvo  
dados\_final[['PTIME']]<-scale(dados\_final[['PTIME']], center =TRUE, scale = TRUE)[,1]  
View(dados\_final)

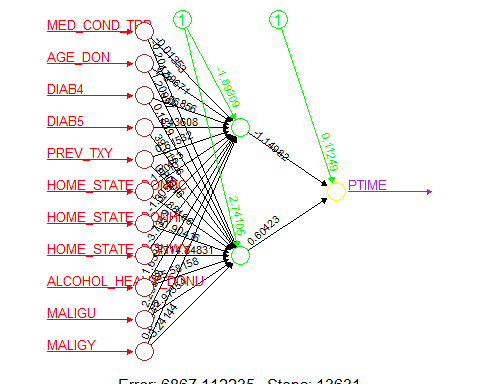
index2 <- sample(1:nrow(dados\_final), dim(dados\_final)[1]\*.70)  
dados\_treino2 <- dados\_final[index2,]  
dados\_teste2 <- dados\_final[-index2,]  
View(dados\_treino2)  
View(dados\_teste2)

**Treinando a Rede neural**

# Colocando em um dataframe  
dados\_treino2 <- data.frame(model.matrix(~ . - 1, data = dados\_treino2))  
dados\_teste2 <- data.frame(model.matrix(~ . - 1, data = dados\_teste2))  
  
modelo\_v2 <- neuralnet::neuralnet(PTIME ~ .,   
 data = dados\_treino2,   
 linear.output = TRUE,  
 hidden = 2,  
 stepmax = 1e7)

**Plotagem do modelo**

# Plot  
plot(modelo\_v2,  
 col.entry.synapse = "red",   
 col.entry = "brown",  
 col.hidden = "green",   
 col.hidden.synapse = "black",  
 col.out = "yellow",   
 col.out.synapse = "purple",  
 col.intercept = "green",   
 fontsize = 10,  
 show.weights = TRUE ,  
 rep = "best")



**Avaliação do modelo**

# Avaliação do modelo  
  
# Avaliação do modelo com dados de treino  
modelo\_v2\_pred\_1 <- compute(modelo\_v2, dados\_treino2)  
rmse\_treino <- rmse(dados\_treino2$PTIME, modelo\_v2\_pred\_1$net.result)  
  
# Avaliação do modelo com dados de teste  
modelo\_v2\_pred\_2 <- compute(modelo\_v2, dados\_teste2)  
rmse\_teste <- rmse(dados\_teste2$PTIME, modelo\_v2\_pred\_2$net.result)  
  
# Exibe os RMSE  
print(paste("RMSE Treino:", rmse\_treino))

## [1] "RMSE Treino: 0.939742669754636"

print(paste("RMSE Teste:", rmse\_teste))

## [1] "RMSE Teste: 0.953911006680501"

Avaliação na unidade original de PTIME

dados\_final2 <- dados\_combinados\_reduzido[, var\_significativas]  
View(dados\_final2)  
  
# Calculando a média de PTIME  
media\_ptime <- mean(dados\_final2$PTIME)  
sd\_ptime <- sd(dados\_final2$PTIME)  
  
#Conversão dos valores  
  
prev\_treino\_originais <- (modelo\_v2\_pred\_1$net.result \* sd\_ptime) + media\_ptime  
prev\_teste\_originais<- (modelo\_v2\_pred\_2$net.result \* sd\_ptime) + media\_ptime  
  
# Calculando RMSE com valores na escala original para dados de treino e teste  
  
  
rmse\_treino\_original <- rmse((dados\_treino2$PTIME \* sd\_ptime) + media\_ptime, prev\_treino\_originais)  
rmse\_teste\_original <- rmse((dados\_teste2$PTIME \* sd\_ptime) + media\_ptime, prev\_teste\_originais)  
  
# Exibe os RMSE na unidade original  
print(paste("RMSE Treino Original:", rmse\_treino\_original))

## [1] "RMSE Treino Original: 311.086888414202"

print(paste("RMSE Teste Original:", rmse\_teste\_original))

## [1] "RMSE Teste Original: 315.777091370956"

1. **– Conclusão**

Conseguimos prever o tempo de sobrevivência dos pacientes 1 ano após receberem um transplante.

O terceiro modelo de regressão linear apresentou a menor taxa de erro (208.848) e, portanto, deve ser usado como versão final. Essa taxa representa cerca de 11% em relação a media de tempo de sobrevivência pós-transplante (1792.6091838).

Para este estudo poderíamos ir mais além para tantar melhorar ainda mais a performance do modelo, trabalhando um pouco mais nos dados, alterando hiperparâmetros ou utilizando outros algorítimos.