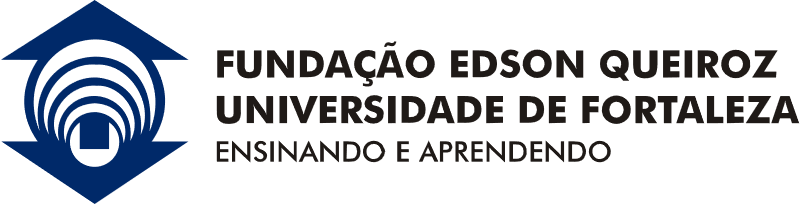
****

**CENTRO DE CIÊNCIAS DA COMUNICAÇÃO E GESTÃO - CCG**

**CURSO: INTELIGÊNCIA DE NEGÓCIOS – EAD**

**DISCIPLINA: ANÁLISE DESCRITIVA DE DADOS**

**DOCENTE ORIENTADOR: JOSE IRAN DA SILVA MELO**

**Atividade final**

Ernandes Alves de Lima Junior – 2313006

Antônio Arlir Rodrigues Dos Santos Filho – 2312990

Jessica Mayara Sousa Oliveira – 2312999

Glaubia Costa Cavalcante de Souza – 2313057

Denise Maria Magalhães Lemos – 2312887

Danilo de Lima Santos – 2313076

## FORTALEZA

## SET/2023

**Introdução**

Neste trabalho são utilizadas as ferramentas R e SQL para explicar e aplicar os conceitos de Análise Descritiva de forma prática, mas não possui o intuito de explicar as ferramentas utilizadas, sendo assim, são apresentados os principais conceitos da Análise Descritiva de Dados e em paralelo a aplicação prática com a linguagem R e a linguagem de consulta SQL. Todo o código está no [arquivo.R](https://github.com/NandesLima/analise-de-dados-com-r-e-sql/blob/main/codigos-dataset/atividade.R).

    Um dos pontos principais é a apresentação e estão sendo utilizadas duas plataformas o [GitHub](https://github.com/NandesLima/analise-de-dados-com-r-e-sql) e o [Youtube](https://www.youtube.com/watch?v=Zf6rcv67bT4).

[Foto editada de uma pessoa

Descrição gerada automaticamente com confiança baixa](https://www.youtube.com/watch?v=Zf6rcv67bT4)

**Requesitos**

* [Interpretador R Versão 4.3.1](https://vps.fmvz.usp.br/CRAN)
* [IDE RStudio versão 2021.09.0](https://posit.co/blog/rstudio-2021.09.0-update-whats-new/)

BIBLIOTECAS UTILIZADAS:

* sqldf – Biblioteca de manipulação de linguagem SQL.
* dplyr – Biblioteca de manipulação de dados.
* corrplot – Biblioteca para matriz de variável de correlação.
* ggplot2 – Biblioteca de criação de gráficos

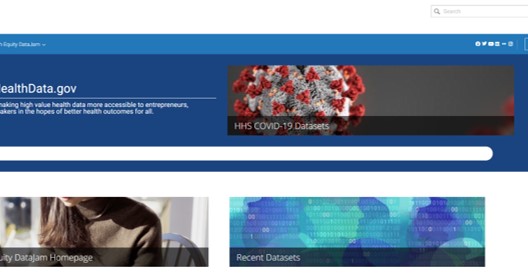
**Dataset**

Usaremos dados de uma pesquisa nacional de custos hospitalares realizada pela US Agency for Healthcare que consiste em registros hospitalares de amostras de pacientes internados. Os dados fornecidos são restritos à cidade de Wisconsin e referem-se a pacientes na faixa etária de 0 a 17 anos.

O [dataset](https://github.com/NandesLima/analise-de-dados-com-r-e-sql/blob/main/codigos-dataset/dataset.csv) foi gerado a partir das seguintes fontes:

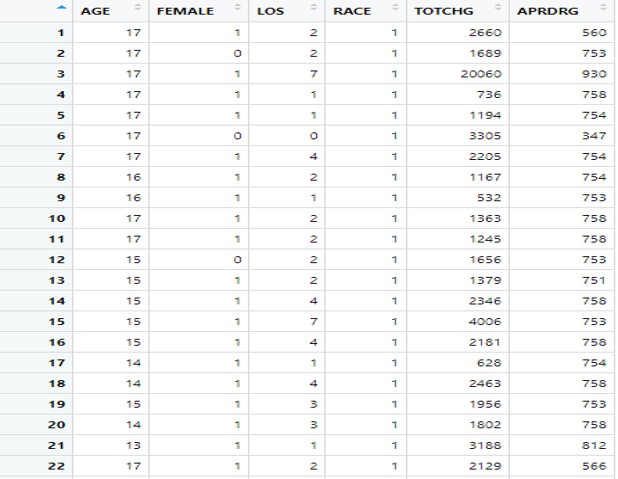
[Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente](https://www.cms.gov/data-research/statistics-trends-and-reports/hospital-cost-report-public-use-file)

[](https://healthdata.gov/)

**Entendendo o Dataset**

* O [dataset](https://github.com/NandesLima/analise-de-dados-com-r-e-sql/blob/main/codigos-dataset/dataset.csv) possui 500 linhas e 6 colunas.
* Todos os atributos são do tipo INT (Número inteiro).



|  |  |
| --- | --- |
| **Atributo** | **Descrição** |
| AGE | Idade do paciente, varia de 0 a 17 anos |
| FEMALE | Variável binária que indica se o paciente é do sexo feminino, sendo 1 para feminino e 0 para masculino |
| LOS | Tempo de internação/permanência do paciente |
| RACE | Raça do paciente, varia de 1 a 6 |
| TOTCHG | Custo de internação |
| APRDRG | Grupo de diagnóstico do paciente |

**Limpeza dos Dados**

Como os dados já estão sumarizados a única limpeza dos dados feita foi excluir os valores nulos. Houve apenas uma ocorrência na coluna RACE.

**Perguntas de Negócios**

Como o intuito não é somente a explicação da Análise Descritiva de Dados, durante cada tópico serão respondidas algumas perguntas de negócios. Elas são:

1. Qual a idade média dos pacientes?
2. Qual o tempo médio de permanência/internações dos pacientes?
3. Qual a moda da idade dos pacientes?
4. Qual é a moda de permanência/internações dos pacientes?
5. Qual a mediana da idade dos pacientes?
6. Qual a mediana do tempo de permanência/internações dos pacientes?
7. Quais as medidas de posição relativa das idades dos pacientes?
8. Quais as medidas de posição relativa dos tempos de permanência/internação dos pacientes?
9. Quais as medidas de dispersão da idade dos pacientes?
10. Quais as medidas de dispersão do tempo de permanência/internação dos pacientes?
11. Qual a distribuição dos pacientes pela raça?
12. Qual a distribuição dos pacientes por idade?
13. Qual o gasto total com internações hospitalares por idade?
14. E qual idade gera o maior gasto total com internações hospitalares?
15. Qual o gasto total com internações hospitalares por gêneros?
16. Qual o gasto médio com internações hospitalares por raça do paciente?
17. Para pacientes acima de 10 anos, qual a média de gastos total com internações hospitalares?
18. Considerando o item anterior, qual idade tem média de gastos superior a 3000?
19. O tempo de permanência é um fator crucial para pacientes internados, é possível descobrir se o tempo de permanência está relacionado com idade, gênero e raça?
20. Quais variáveis têm maior impacto nos custos de internação hospitalar?

**Análise Descritiva de Dados**

A Análise Descritiva de Dados é responsável pela coleta, organização, descrição, síntese e análise dos dados. A Análise Descritiva pode ser feita através de MEDIDAS DE TENDÊNCIA CENTRAL, DE POSIÇÃO RELATIVA, DE DISPERSÃO, além de TABELAS DE FREQUÊNCIA E REGRESSÃO.

**Medidas de Tendência Central**

**Medidas de tendência central** são aqueles que mostram o comportamento dos dados em torno de uma medida de centro, temos a média, moda e mediana.

#### [Média](https://github.com/NandesLima/analise-de-dados-com-r-e-sql#m%C3%A9dia)

**Média** é a mais utilizada das medidas de tendência central, é o resultado das somas dos valores de uma variável dividido pela quantidade de observações.

O conceito de média responde as duas primeiras perguntas e no código atividade. R é utilizado a linguagem R em paralelo com o SQL.

1. Qual a idade média dos pacientes?

R: 5,096192 anos

1. Qual o tempo médio de permanência/internações dos pacientes?

R: 2,829659 horas

#### [Moda](https://github.com/NandesLima/analise-de-dados-com-r-e-sql#moda)

**Moda** é o valor que se repete mais vezes dentre os dados observados.

E com ela respondemos as questões 3 e 4 de negócios, utilizando novamente em paralelo o R e o SQL.

1. Qual é a moda da idade dos pacientes?

R: 0 anos

1. Qual é a moda de permanência/internações dos pacientes?

R: 2 horas

#### [Mediana](https://github.com/NandesLima/analise-de-dados-com-r-e-sql#mediana)

**Mediana** é o valor do meio do conjunto de dados, organizado de forma crescente ou decrescente.

A mediana responde as perguntas 5 e 6.

1. Qual a mediana da idade dos pacientes?

R: 0 anos

1. Qual a mediana do tempo de permanência/internações dos pacientes?

R: 2 horas

**Medidas de Posição Relativa**

**Medidas de posição relativa** comparam a posição de um valor em relação ao de outro valor em um conjunto de dados. Percentis e quartis sãos os mais comuns de serem utilizados.

* **Percentis** dividem o conjunto de dados em 100 partes iguais.
* **Quartis** dividem o conjunto de dados em 4 partes iguais.

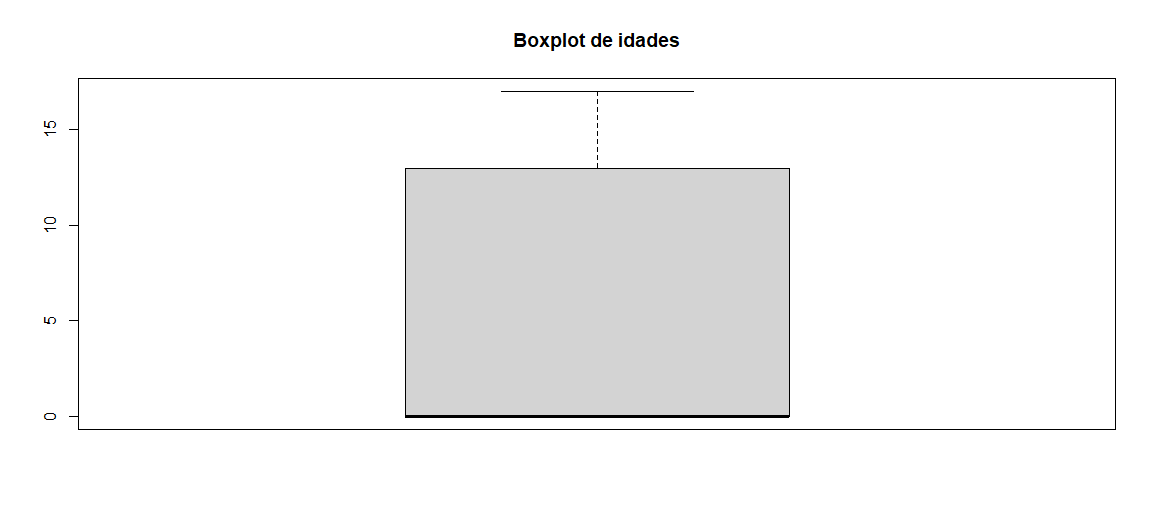
Quartis dividem o conjunto de dados em 25% (primeiro quartil), 50% (segundo quartil), 75% (terceiro quartil) e 100% (quarto quartil).

O segundo quartil ou o 50% percentil é a mediana.

As medidas de posição relativa respondem as perguntas 7 e 8 e é utilizado apenas a linguagem R, e junto com as respostas, das medidas de posição é apresentado em seguida um gráfico de boxplot que mostra o comportamento das medidas de posição relativa.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Min.** | **1st Qu.** | **Median** | **Mean** | **3rd Qu.** | **Max.** |
| 0,000 | 0,000 | 0,000 | 5,096 | 13,000 | 17,000 |

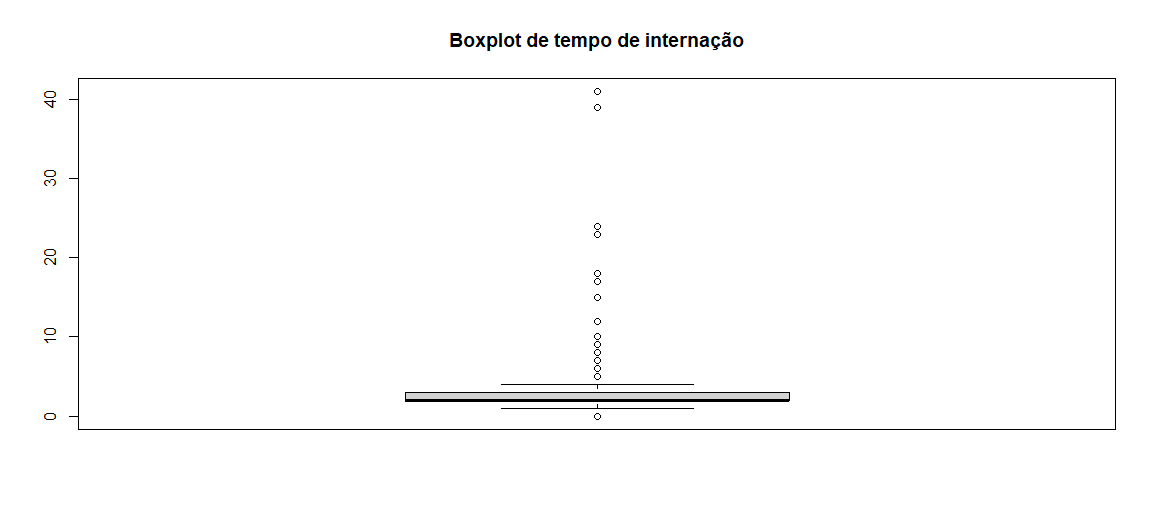
1. Quais as medidas de posição relativa das idades dos pacientes?



**R**: Analisando os valores tanto das medidas de posição relativa como do gráfico bloxpot, podemos concluir que pelo menos 50% dos dados de idade estão na faixa de 0 anos, ou seja, recém-nascidos.

1. Quais as medidas de posição relativa dos tempos de permanência/internação dos pacientes?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Min.** | **1st Qu.** | **Median** | **Mean** | **3rd Qu.** | **Max.** |
| 0,00 | 2,00 | 2,00 | 2,83 | 3,00 | 41,00 |



**R**: Os dados que estão concentrados entre o primeiro quartil e a mediana que é 25% dos dados são referentes a 2 horas de permanência/internação no hospital. Observamos também muitos outliers, tanto o valor 0 como vários após as 3 horas.

**Medidas de Dispersão**

**Medidas de dispersão** mostram o quão os valores estão espalhados dentro do conjunto de dados: Variância, Desvio Padrão e Coeficiente de variação (CV).

* **Variância** mede a variabilidade dos dados em relação à média.
* **Desvio padrão** em termos simples é a distância média que os valores têm da média, ou seja, como seu nome sugere é um padrão de desvio (distância) em relação à média.
* **Coeficiente de variação (CV)** mede o desvio padrão em termos percentuais em relação da média, quanto maior o CV maior a variabilidade dos dados e menor sua consistência e quanto menor, menor é sua variabilidade e maior a consistência dos dados.

1. Quais as medidas de dispersão da idade dos pacientes?

|  |  |  |
| --- | --- | --- |
| **Variância** | **Desvio Padrão** | **CV** |
| 48,34013 | 6,952706 | 136,4294% |

**R**: As medidas de dispersão permitem uma interpretação de como os dados estão espalhados em relação a média, mas é necessário comparar com outro grupo de dados. Pelo menos com o CV podemos ver que o valor de espalhamento dos dados está em 136% em relação à média, logo podemos ver que os dados são muito dispersos.

|  |  |  |
| --- | --- | --- |
| **Variância** | **Desvio Padrão** | **CV** |
| 11,33438 | 3,366657 | 118,9775% |

1. Quais as medidas de dispersão do tempo de permanência/internação dos pacientes?

**R**: Pelo CV podemos perceber que o valor é 118% de dispersão em relação a média, ou seja, os dados são muito dispersos dentro do conjunto de dados.

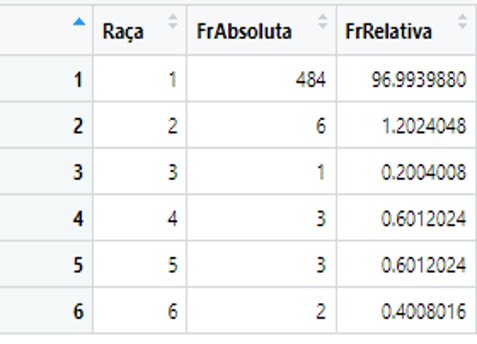
**Tabelas de Frequência**

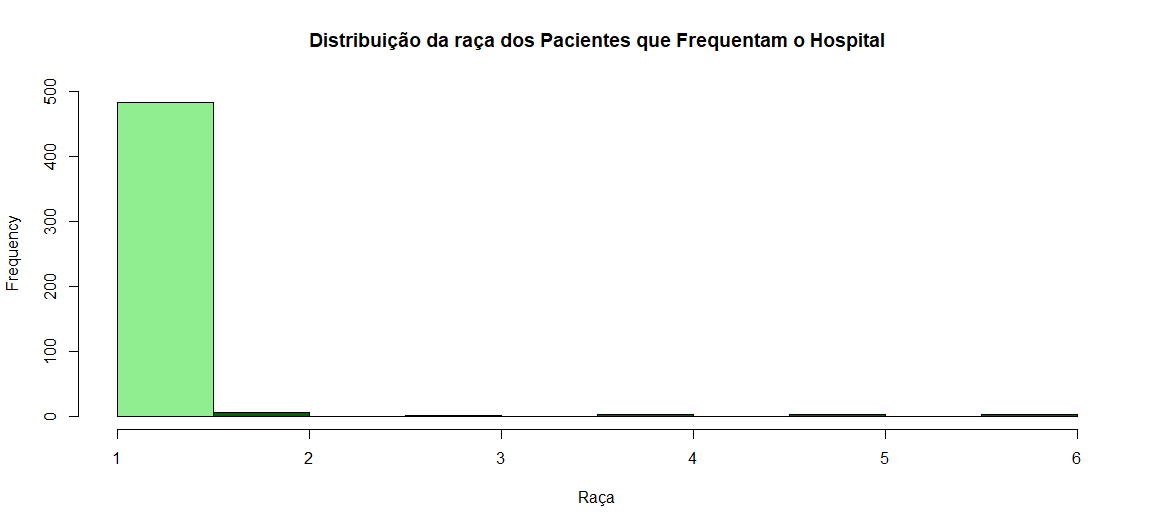
**As tabelas de frequência** sumarizam a quantidade observada de determinado atributo, ou seja, a frequência que aparece nos dados.

* **Frequência absoluta** número de eventos observados.
* **Frequência relativa** relação entre os eventos observados e o total, ou seja, pode ser dada como uma fração ou em porcentagem.

Aqui utilizaremos as tabelas de frequência utilizando a linguagem R e SQL para responder as perguntas 11 e 12, além de gerar gráficos de histograma.

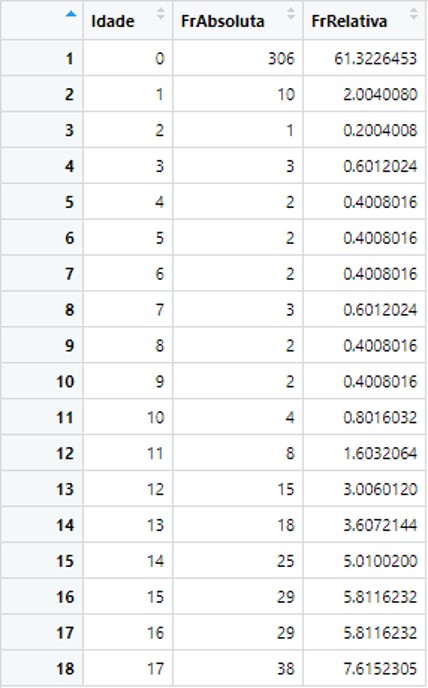
1. Qual a distribuição dos pacientes pela raça?

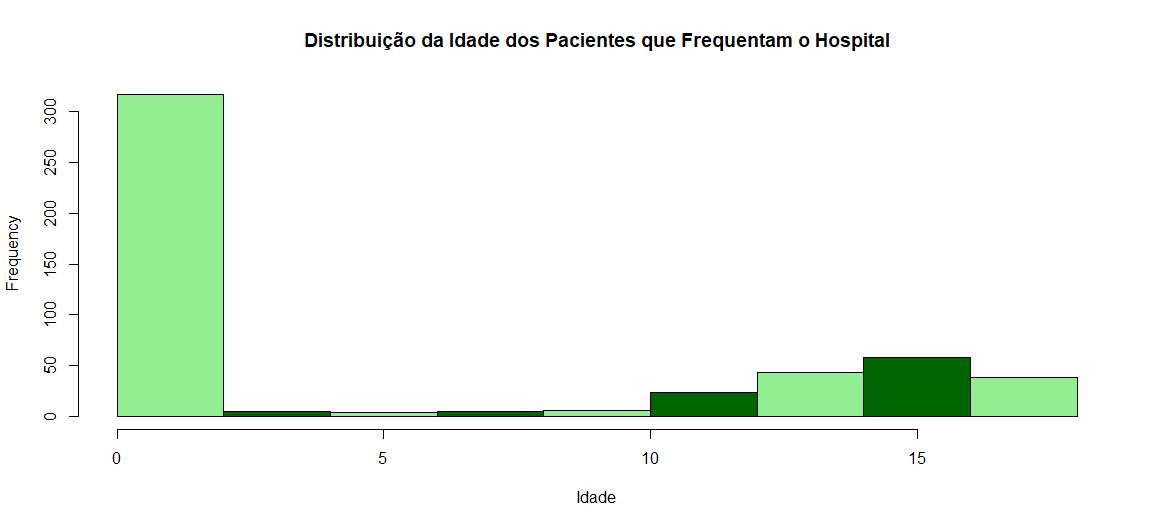




**R:** Pode-se concluir que pela distribuição a raça mais atendida é a raça de número 1 que se repete 464 vezes, um total de 96% dos dados.

1. Qual a distribuição dos pacientes por idade?



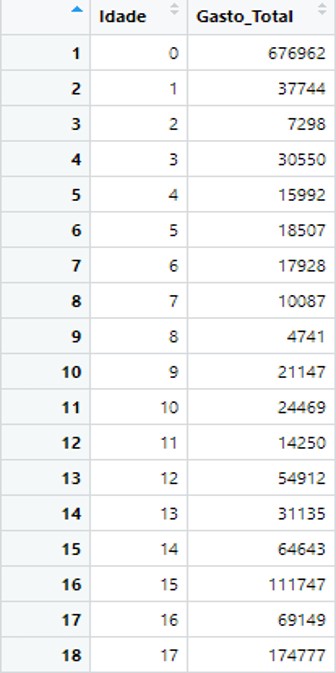


**R:** A idade que mais se repete é a de 0 anos, recém-nascidos, 306 vezes, um total de

**Análise Exploratória**

Aqui respondemos algumas perguntas de negócio (13 a 18) com algumas sumarizações de dados.

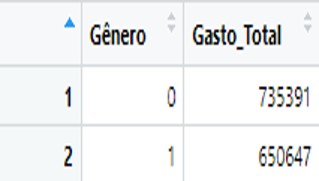
1. Qual o gasto total com internações hospitalares por idade?



1. E qual idade gera o maior gasto total com internações hospitalares?

|  |  |
| --- | --- |
| **Idade** | **Gasto total** |
| 0 | US$ 676.952,00 |

1. Qual o gasto total com internações hospitalares por gêneros?

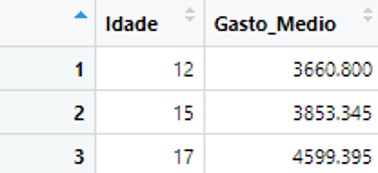


1. Qual o gasto médio com internações hospitalares por raça do paciente?



1. Para pacientes acima de 10 anos, qual a média de gastos total com internações hospitalares?



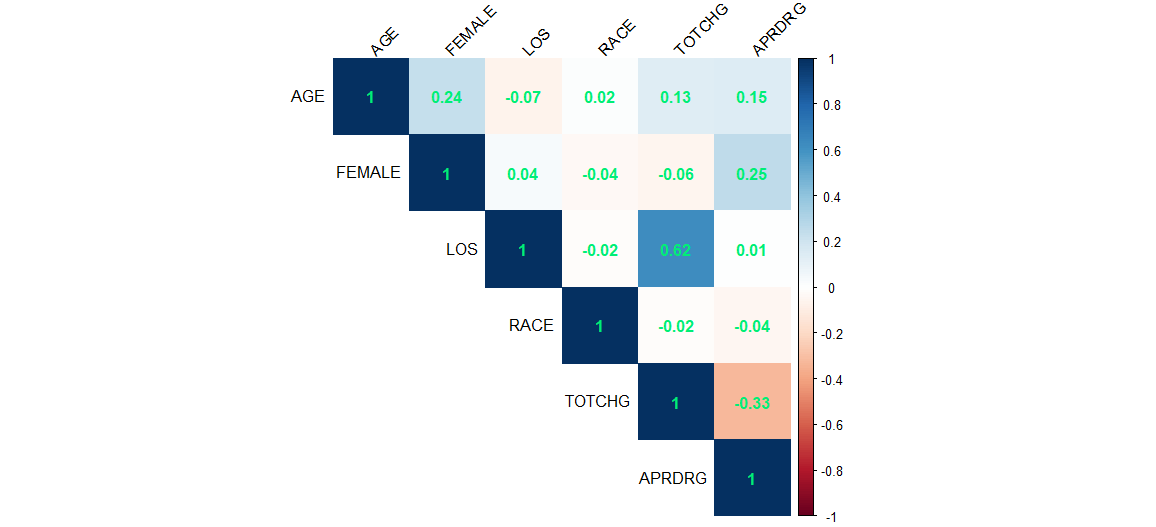
1. Considerando o item anterior, qual idade tem média de gastos superior a 3000?

**Regressão Linear**

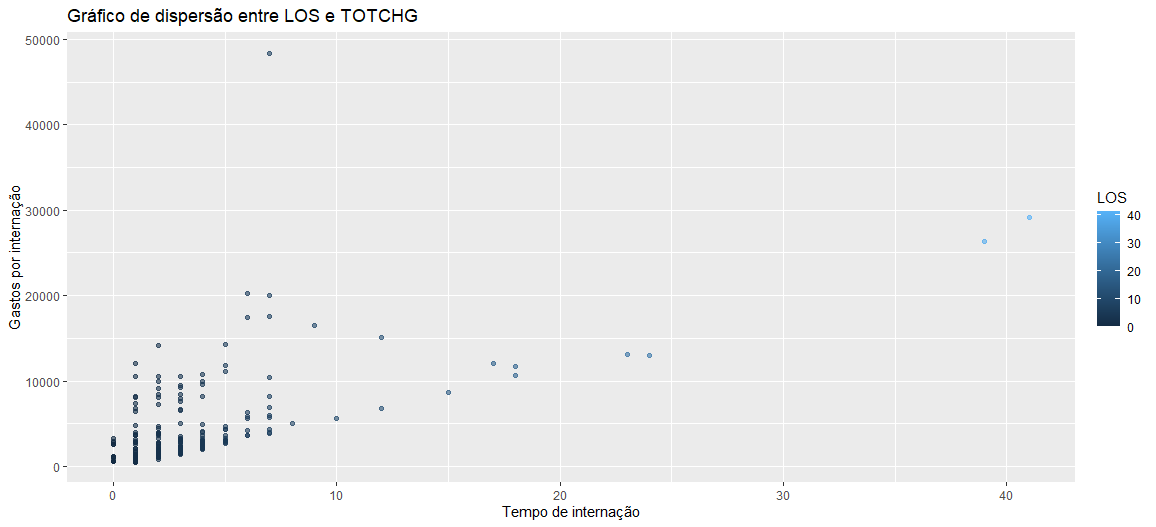
O modelo de regressão investiga a relação entre variáveis.

### [Regressão linear simples](https://github.com/NandesLima/analise-de-dados-com-r-e-sql#regress%C3%A3o-linear-simples)

* **Regressão linear simples** descreve o relacionamento entre duas variáveis usando uma equação, uma variável independente x que explica a variação em outra variável, que é chamada de variável dependente y. O gráfico de dispersão entre as duas variáveis forma uma linha reta.
* A regressão linear utiliza-se da correlação, que é a força de relacionamento linear entre as variáveis.
* O coeficiente de correlação r permite identificar se o relacionamento entre duas variáveis é forte ou não o suficiente para considerar estatisticamente significante.
* O coeficiente r varia entre -1 (forte relação negativa) e 1 (forte relação positiva), sendo 0 sem relação.
* Correlação não significa causalidade (um evento causado pelo outro).



A matriz gerada mostra o conficiênte de correlação entre os dados do dataset, e podemos verificar que o único relacionamento estatisticamente significante é entre o tempo de internação (LOS) e os custos (TOTCHG).

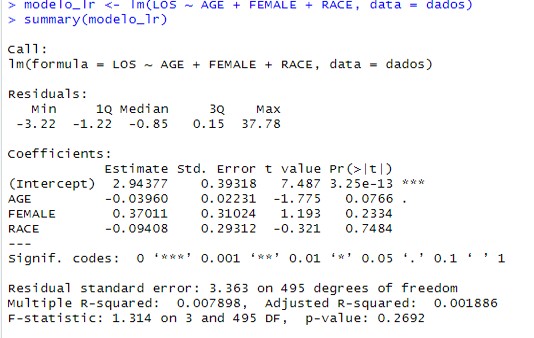


Podemos observar que o gráfico tem uma tendência linear, mais com uma certa dispersão relevante.

**Regressão Linear Múltipla**

**Regressão linear múltipla** é a regressão linear que possui uma variável dependente y e duas ou mais variáveis independentes x1...xn.

1. O tempo de permanência é um fator crucial para pacientes internados, é possível descobrir se o tempo de permanência está relacionado com idade, gênero e raça?

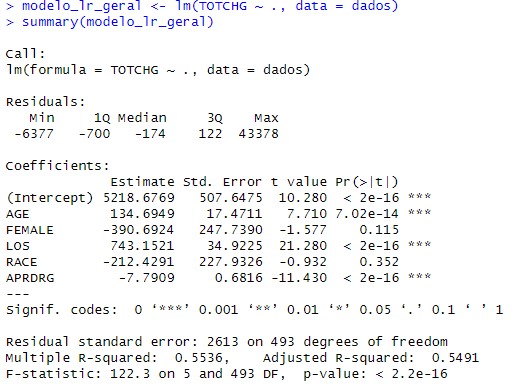


Pr(> |T|) é o Valor-p e no mercado é utilizado normalmente com significância menor que 0,05 para ter relevância estatística, logo esse modelo de linearidade falha, pois todos os valores em p são maiores que 0,05.

O coeficiente de correlação também está muito próximo de 0, o que prova não haver uma correlação significativa entre as variáveis.

A quantidade de asteriscos no relatório indica a significância, sendo 3 asteriscos alta significância, logo como as variáveis independentes não possuem asteriscos não possuem significância estatística.

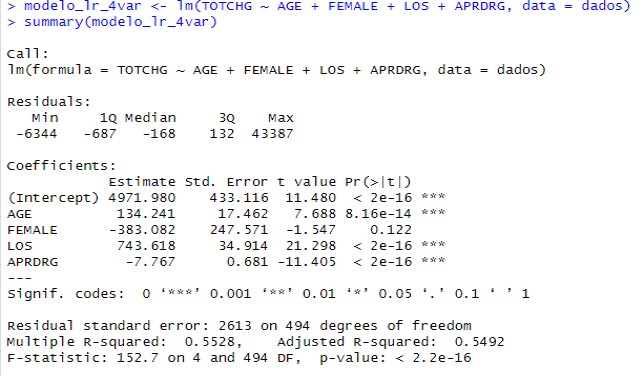
1. Quais variáveis têm maior impacto nos custos de internação hospitalar?



* **Modelo de teste 1**

Algumas variáveis dependentes passam no teste do p-value: AGE, LOS e APRDRG.

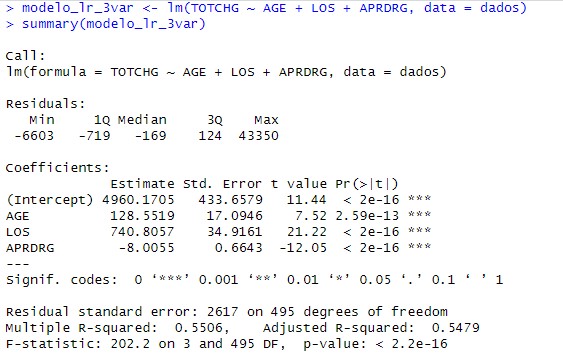
As variáveis FEMALE e RACE não passam então será feito um novo teste, será removida a variável RACE que teve o menor impacto.



* **Modelo de teste 2**

As variáveis dependentes que passaram no teste anterior do p-value, passaram novamente: AGE, LOS e APRDRG.

A variável FEMALE não passa novamente e será removida para o próximo modelo.



* **Modelo de teste 3**

Todas as variáveis dependentes passam no teste p-value mostrando uma alta significância.

**Conclusão:** Os custos dos cuidados de saúde, dependem da idade, do tempo de permanência e do grupo de diagnóstico.

**Conclusão**

A análise descritiva de dados permite o entendimento do conjunto de dados desde o seu centro e dispersão em torno do centro. Sumarizações como nas tabelas de frequência. E até mesmo entender a relação entre os dados através dos processos de regressão. Sendo úteis até para responder questões práticas de negócio.

Verificamos também que no vídeo postado no Youtube tivemos comentários positivos, tanto no elogio em relação ao trabalho, como nos insights do uso do R e SQL para responder as perguntas de negócio.