FUNDAÇÃO EDSON QUEIROZ
UNIVERSIDADE DE FORTALEZA
ENSINANDO E APRENDENDO

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.

Um dos pontos principais é a apresentação e estão sendo utilizadas duas plataformas o GitHub e o Youtube.

Requesitos

- Interpretador R Versão 4.3.1
- IDE RStudio versão 2021.09.0

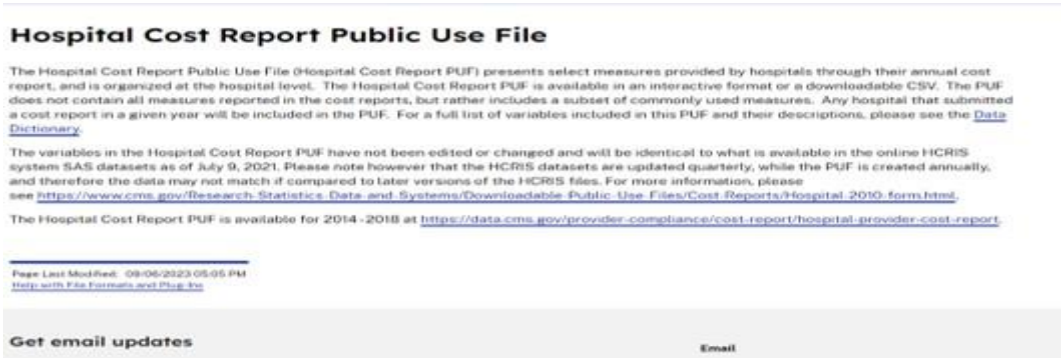
BIBLIOTECAS UTILIZADAS:

- 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 foi gerado a partir das seguintes fontes:



Hospital Cost Report Public Use File

The Hospital Cost Report Public Use File (Hospital Cost Report PUF) presents select measures provided by hospitals through their annual cost report, and is organized at the hospital level. The Hospital Cost Report PUF is available in an interactive format or a downloadable CSV. The PUF does not contain all measures reported in the cost reports, but rather includes a subset of commonly used measures. Any hospital that submitted a cost report in a given year will be included in the PUF. For a full list of variables included in this PUF and their descriptions, please see the Data Dictionary.

The variables in the Hospital Cost Report PUF have not been edited or changed and will be identical to what is available in the online HCRIS system SAS datasets as of July 9, 2021. Please note however that the HCRIS datasets are updated quarterly, while the PUF is created annually, and therefore the data may not match if compared to later versions of the HCRIS files. For more information, please see https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/Cost-Reports/Hospital-2010-form.html.

The Hospital Cost Report PUF is available for 2014 - 2018 at https://data.cms.gov/provider-compliance/cost-report/hospital-provider-cost-report.

Page Last Modified: 09/06/2023 05:05 PM
Help with File Formats and Plug-Ins

Get email updates

Email

Entendendo o Dataset

- O dataset possui 500 linhas e 6 colunas.
- Todos os atributos são do tipo INT (Número inteiro).

	AGE	FEMALE	LOS	RACE	TOTCHG	APRDRG
1	17	1	2	1	2660	560
2	17	0	2	1	1689	753
3	17	1	7	1	20060	930
4	17	1	1	1	736	758
5	17	1	1	1	1194	754
6	17	0	0	1	3305	347
7	17	1	4	1	2205	754
8	16	1	2	1	1167	754
9	16	1	1	1	532	753
10	17	1	2	1	1363	758
11	17	1	2	1	1245	758
12	15	0	2	1	1656	753
13	15	1	2	1	1379	751
14	15	1	4	1	2346	758
15	15	1	7	1	4006	753
16	15	1	4	1	2181	758
17	14	1	1	1	628	754
18	14	1	4	1	2463	758
19	15	1	3	1	1956	753
20	14	1	3	1	1802	758
21	13	1	1	1	3188	812
22	17	1	2	1	2129	566

Atri-buto	Descrição
AGE	Idade do paciente, varia de 0 a 17 anos
FE-MALE	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
TO-TCHG	Custo de internação
APR-DRG	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

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

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

R: 2,829659 horas

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.

3. Qual é a moda da idade dos pacientes?

R: 0 anos

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

R: 2 horas

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.

5. Qual a mediana da idade dos pacientes?

R: 0 anos

6. 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.

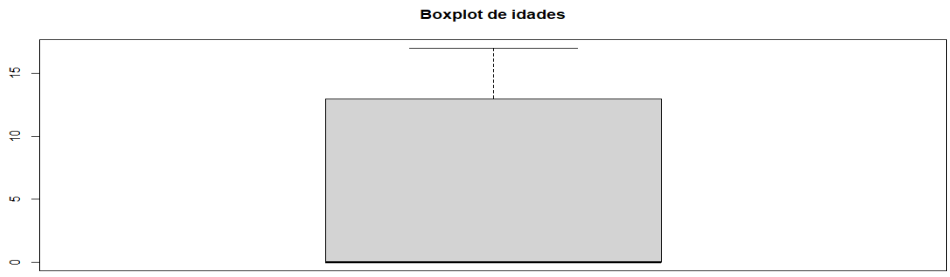
Quartis dividem o conjunto de dados em 25% (primeiro quartil), 50% (segundo quartil), 75% (terceiro quartil) e 100% (quarto quartil).

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.

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

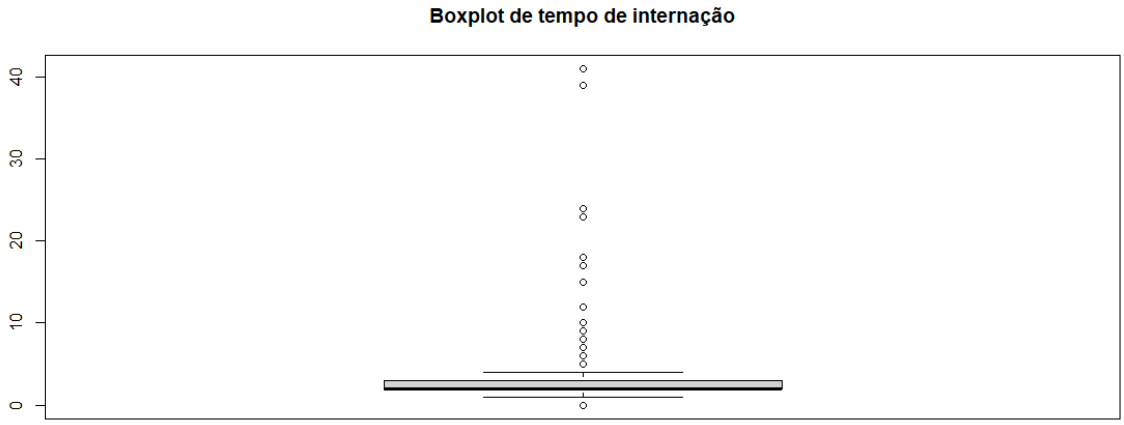
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0,000	0,000	0,000	5,096	13,000	17,000



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.

8. 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.

9. 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.

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

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

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

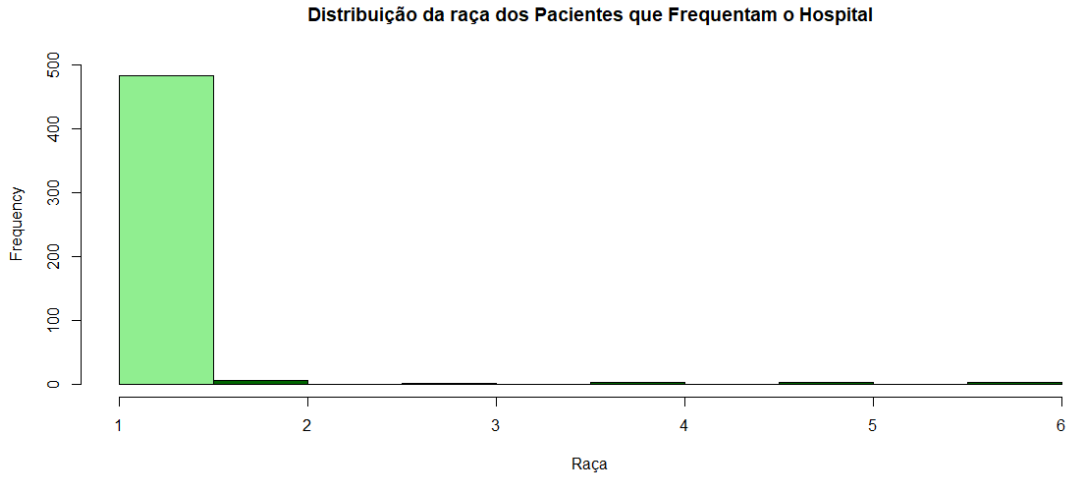
As tabelas de frequência sumarizam a quantidade observada de determinado atributo, ou seja, a frequência que aparece nos dados.

- **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.

11. Qual a distribuição dos pacientes pela raça?

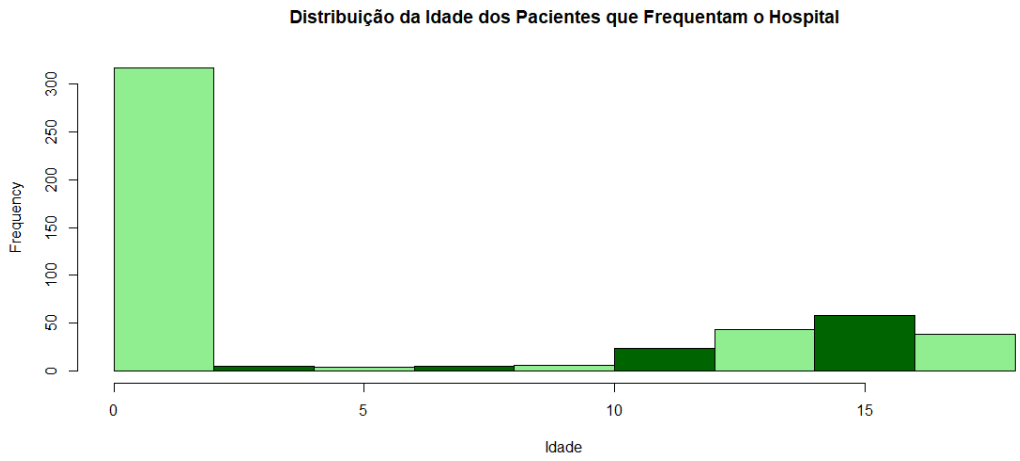
	Raça	FrAbsoluta	FrRelativa
1	1	484	96.9939880
2	2	6	1.2024048
3	3	1	0.2004008
4	4	3	0.6012024
5	5	3	0.6012024
6	6	2	0.4008016



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.

12. Qual a distribuição dos pacientes por idade?

	Idade	FrAbsoluta	FrRelativa
1	0	306	61.3226453
2	1	10	2.0040080
3	2	1	0.2004008
4	3	3	0.6012024
5	4	2	0.4008016
6	5	2	0.4008016
7	6	2	0.4008016
8	7	3	0.6012024
9	8	2	0.4008016
10	9	2	0.4008016
11	10	4	0.8016032
12	11	8	1.6032064
13	12	15	3.0060120
14	13	18	3.6072144
15	14	25	5.0100200
16	15	29	5.8116232
17	16	29	5.8116232
18	17	38	7.6152305



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.

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

	Idade	Gasto_Total
1	0	676962
2	1	37744
3	2	7298
4	3	30550
5	4	15992
6	5	18507
7	6	17928
8	7	10087
9	8	4741
10	9	21147
11	10	24469
12	11	14250
13	12	54912
14	13	31135
15	14	64643
16	15	111747
17	16	69149
18	17	174777

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

Idade	Gasto total
0	US$ 676.952,00

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

	Gênero	Gasto_Total
1	0	735391
2	1	650647

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

	Raça	Gasto_Medio
1	1	2772.669
2	2	4202.167
3	3	3041.000
4	4	2344.667
5	5	2026.667
6	6	1349.000

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

	Idade	Gasto_Medio
1	11	1781.250
2	12	3660.800
3	13	1729.722
4	14	2585.720
5	15	3853.345
6	16	2384.448
7	17	4599.395

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

	Idade	Gasto_Medio
1	12	3660.800
2	15	3853.345
3	17	4599.395

Regressão Linear

O modelo de regressão investiga a relação entre variáveis.

Regressão linear simples

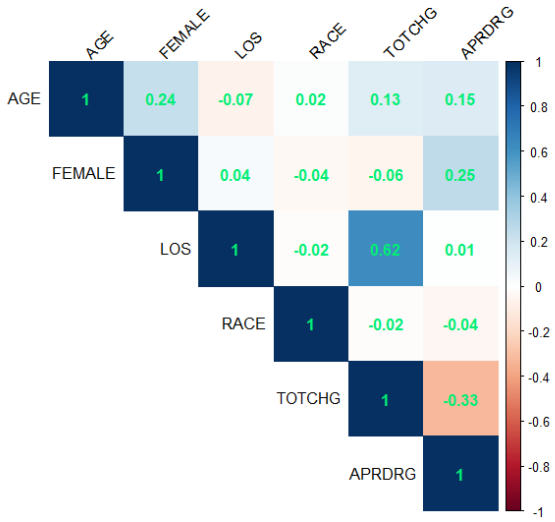
- **Regressão linear simples** descreve o relacionamento entre duas variáveis usando uma equação, uma variável independente x que explica a variação em outra variável, que é chamada de variável dependente y. O gráfico de dispersão entre as duas variáveis forma uma linha reta.

- A regressão linear utiliza-se da correlação, que é a força de relacionamento linear entre as variáveis.

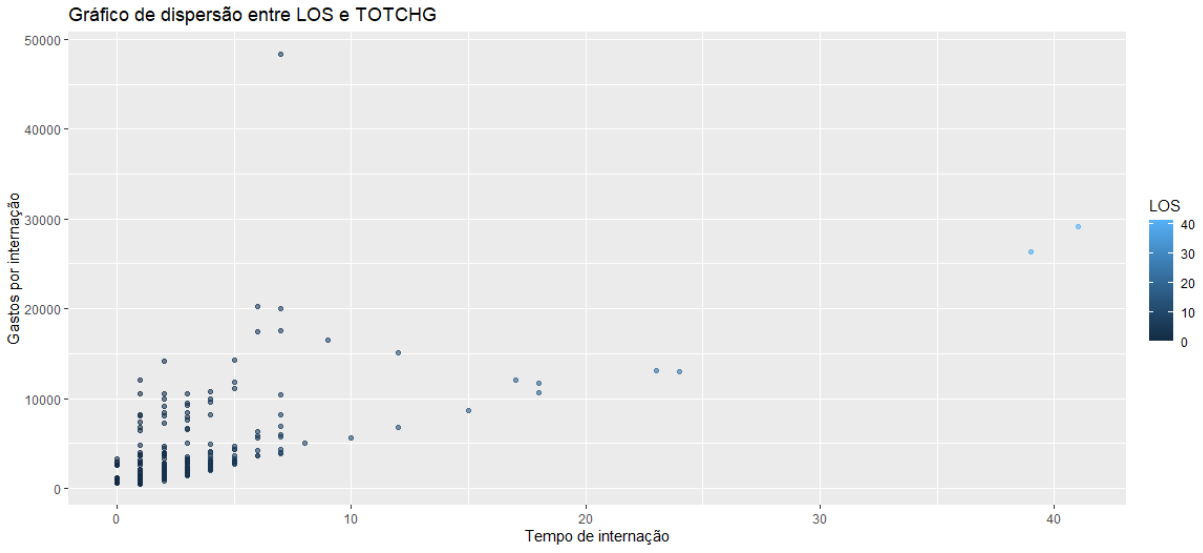
- O coeficiente de correlação r permite identificar se o relacionamento entre duas variáveis é forte ou não o suficiente para considerar estatisticamente significante.

- 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.

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?

```
> modelo_lr <- lm(LOS ~ AGE + FEMALE + RACE, data = dados)
> summary(modelo_lr)

Call:
lm(formula = LOS ~ AGE + FEMALE + RACE, data = dados)

Residuals:
   Min     1Q Median     3Q    Max
 -3.22  -1.22  -0.85   0.15  37.78

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.94377    0.39318   7.487 3.25e-13 ***
AGE         -0.03960    0.02231  -1.775   0.0766 .
FEMALE       0.37011    0.31024   1.193   0.2334
RACE        -0.09408    0.29312  -0.321   0.7484
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.363 on 495 degrees of freedom
Multiple R-squared:  0.007898,  Adjusted R-squared:  0.001886
F-statistic: 1.314 on 3 and 495 DF,  p-value: 0.2692
```

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.

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

```
> modelo_lr_geral <- lm(TOTCHG ~ ., data = dados)
> summary(modelo_lr_geral)

Call:
lm(formula = TOTCHG ~ ., data = dados)

Residuals:
   Min     1Q Median     3Q    Max
 -6377   -700   -174    122  43378

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5218.6769   507.6475  10.280  < 2e-16 ***
AGE          134.6949    17.4711   7.710 7.02e-14 ***
FEMALE      -390.6924   247.7390  -1.577    0.115
LOS          743.1521    34.9225  21.280  < 2e-16 ***
RACE        -212.4291   227.9326  -0.932    0.352
APRDRG        -7.7909     0.6816 -11.430  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2613 on 493 degrees of freedom
Multiple R-squared:  0.5536,    Adjusted R-squared:  0.5491
F-statistic: 122.3 on 5 and 493 DF,  p-value: < 2.2e-16
```

- **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_lr_4var <- lm(TOTCHG ~ AGE + FEMALE + LOS + APRDRG, data = dados)
> summary(modelo_lr_4var)

Call:
lm(formula = TOTCHG ~ AGE + FEMALE + LOS + APRDRG, data = dados)

Residuals:
   Min     1Q Median     3Q    Max
 -6344   -687   -168    132  43387

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4971.980    433.116  11.480  < 2e-16 ***
AGE          134.241     17.462   7.688 8.16e-14 ***
FEMALE      -383.082    247.571  -1.547    0.122
LOS          743.618     34.914  21.298  < 2e-16 ***
APRDRG        -7.767      0.681 -11.405  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2613 on 494 degrees of freedom
Multiple R-squared:  0.5528,    Adjusted R-squared:  0.5492
F-statistic: 152.7 on 4 and 494 DF,  p-value: < 2.2e-16
```

- **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_lr_3var <- lm(TOTCHG ~ AGE + LOS + APRDRG, data = dados)
> summary(modelo_lr_3var)

Call:
lm(formula = TOTCHG ~ AGE + LOS + APRDRG, data = dados)

Residuals:
   Min     1Q Median     3Q    Max
 -6603   -719   -169    124  43350

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4960.1705   433.6579   11.44  < 2e-16 ***
AGE          128.5519    17.0946    7.52 2.59e-13 ***
LOS          740.8057    34.9161   21.22  < 2e-16 ***
APRDRG        -8.0055     0.6643  -12.05  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2617 on 495 degrees of freedom
Multiple R-squared:  0.5506,    Adjusted R-squared:  0.5479
F-statistic: 202.2 on 3 and 495 DF,  p-value: < 2.2e-16
```

- **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.