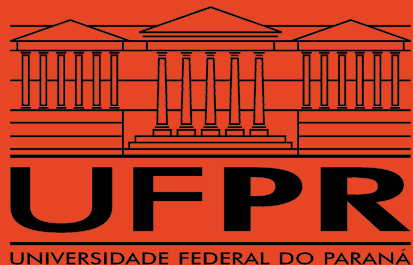


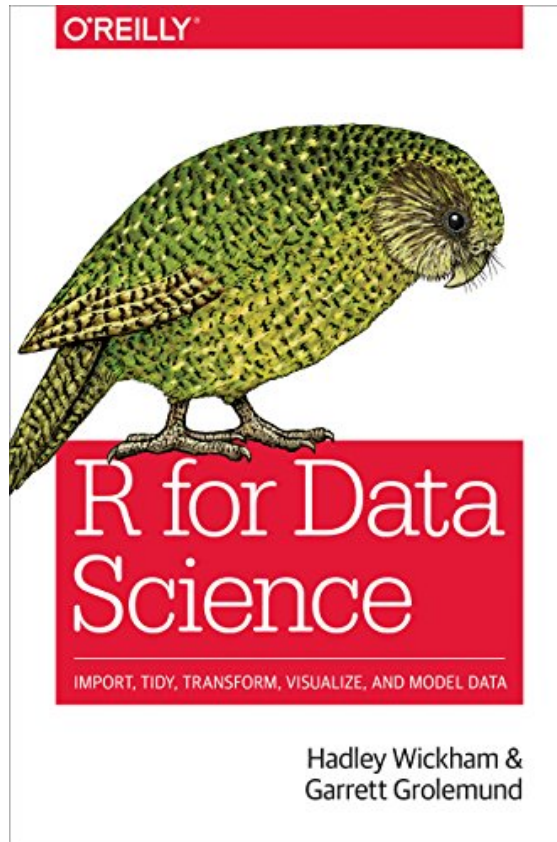
Conhecendo o Tidyverse

stringr, forcats e lubridate

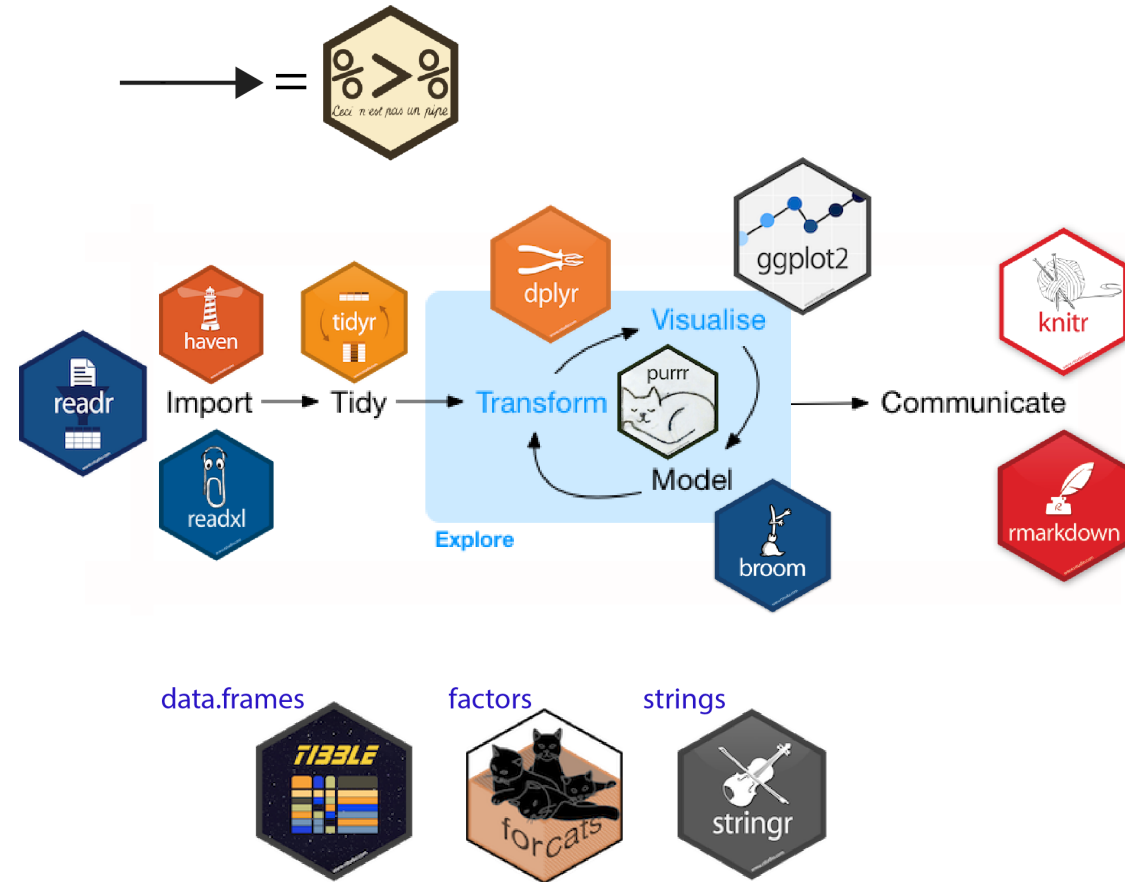
Fernando de Pol Mayer (LEG/DEST/UFPR)
2022-03-29



R for Data Science



R for Data Science, a principal referência sobre o emprego da linguagem R em ciência de dados.



Workflow de ciência de dados com o {tidyverse}. Fonte:
https://oliviergimenez.github.io/intro_tidyverse/#7

{stringr}

O que são expressões regulares?

Usos de regex

- ▶ Padrões em cadeias de caracteres.
- ▶ Detectar.
- ▶ Extrair.
- ▶ Eliminar.
- ▶ Substituir.

Exemplos:

- ▶ Telefone.
- ▶ Email.
- ▶ Data.
- ▶ CPF/CNPJ.
- ▶ Endereço.

Opinião do Dono

dados semi-estruturados

"Melhor hatch pequeno"
Volkswagen Gol G7 Track 1.0 2017/2017 ← *categórica*
Renato - Campo Grande MS ← *geográfica*
Dono há 1 ano - 11.090 km ← *duração*
Carro anterior: Ford Ecosport ← *transação*

dados não estruturados

Prós:
Aparência externa muito bonita, moderna, interior show, igual ao do Golf, painel completo, bonito, com computador de bordo, multimídia, rodas de liga leve, controles no volante, econômico, desempenho surpreendente, nem parece 1.0. A grade dianteira dele, aliada aos faróis de design moderno, dão um ar mais agressivo e esportivo ao carro, deixando ele muito atraente.

Contras:
Porta malas pequeno, espaço para quem viaja atrás, no meio, bastante desconfortável, ruídos nas portas, alguns ruídos o painel, suspensão dura.

Defeitos apresentados:
Até agora, somente caiu o chip da chave canivete, que impossibilitava o funcionamento do carro.

Opinião Geral:
A aparência desse Gol me surpreendeu positivamente, um carro muito bonito, sobretudo o interior. Quem estava acostumado com o Gol G5, vai se surpreender também. O motor de 3 cilindros, a princípio me deixou um tanto preocupado, mas o desempenho é excelente, bastante ágil, além de deixa-lo muito econômico. O preço é salgado, paguei R\$ 45.000,00, contudo, a revenda é garantida, pois o Gol ainda é um dos carros usados mais procurados no mercado. Boa compra, eu recomendo.

11/02/2018 23:25:00 ← *cronológica*

dados estruturados

Estilo	★★★★★	10
Acabamento	★★★★☆	7
Posição de dirigir	★★★★☆	7
Instrumentos	★★★★★	10
Interior	★★★☆☆	6
Porta-malas	★★☆☆☆	5
Desempenho	★★★★★	9
Motor	★★★★★	8
Câmbio	★★★★★	9
Freios	★★★★★	8
Suspensão	★★★☆☆	6
Consumo	★★★★★	10
Estabilidade	★★★★★	8
Custo-Benefício	★★★★★	8
Recomendação	★★★★★	8
Avaliação Geral	★★★★★	7,93

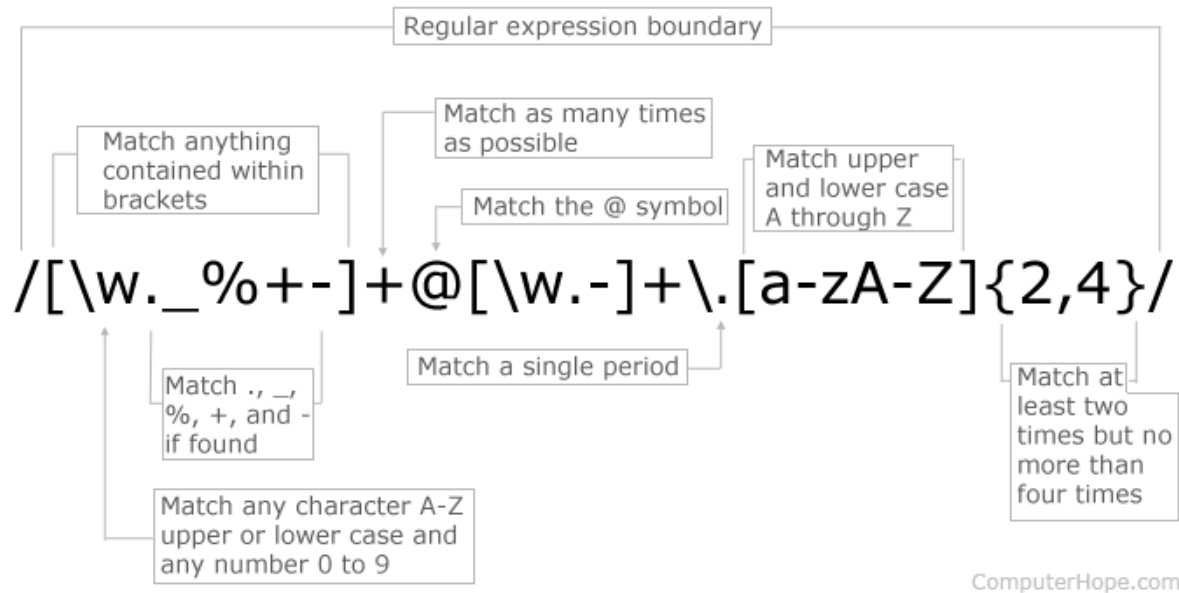
discreta limitada

Avaliação online de um veículo do site carros na web.

Fonte: <https://github.com/leg-ufpr/hackathon/blob/master/datasets.md>.

Expressões regulares

Regular Expression E-mail Matching Example



Regex online

- ▶ <https://regexr.com/>.
- ▶ <https://regex101.com/>.
- ▶ <https://www.regextester.com/>.
- ▶ <https://regex-generator.olafneumann.org/>.

Expressão regular para verificar emails.

Fonte: <https://paulvanderlaken.com/2017/10/03/regular-expressions-in-r-part-1-introduction-and-base-r-functions/>.

Conteúdo do pacote

- ▶ Página do pacote: <https://stringr.tidyverse.org/>.
- ▶ Vinheta de uso: <https://stringr.tidyverse.org/articles/regular-expressions.html>.
- ▶ Cheatsheet: <https://github.com/rstudio/cheatsheets/raw/master/strings.pdf>

Principais funções

- ▶ `str_detect()`: detectar padrões para usar em filtros.
- ▶ `str_remove*()`: remove um padrão.
- ▶ `str_extract*()`: extrai um padrão.
- ▶ `str_replace*()`: Substitui um padrão.
- ▶ `str_c()`: concatena strings.

A maioria das funções do pacote tem o prefixo `str_`.

<code>%>%</code>	<code>str_glue_data</code>	<code>str_starts</code>
<code>boundary</code>	<code>str_interp</code>	<code>str_sub</code>
<code>coll</code>	<code>str_length</code>	<code>str_sub<-</code>
<code>fixed</code>	<code>str_locate</code>	<code>str_subset</code>
<code>fruit</code>	<code>str_locate_all</code>	<code>str_to_lower</code>
<code>invert_match</code>	<code>str_match</code>	<code>str_to_sentence</code>
<code>regex</code>	<code>str_match_all</code>	<code>str_to_title</code>
<code>sentences</code>	<code>str_order</code>	<code>str_to_upper</code>
<code>str_c</code>	<code>str_pad</code>	<code>str_trim</code>
<code>str_conv</code>	<code>str_remove</code>	<code>str_trunc</code>
<code>str_count</code>	<code>str_remove_all</code>	<code>str_view</code>
<code>str_detect</code>	<code>str_replace</code>	<code>str_view_all</code>
<code>str_dup</code>	<code>str_replace_all</code>	<code>str_which</code>
<code>str_ends</code>	<code>str_replace_na</code>	<code>str_wrap</code>
<code>str_extract</code>	<code>str_sort</code>	<code>word</code>
<code>str_extract_all</code>	<code>str_split</code>	<code>words</code>
<code>str_flatten</code>	<code>str_split_fixed</code>	
<code>str_glue</code>	<code>str_squish</code>	

{forcats}

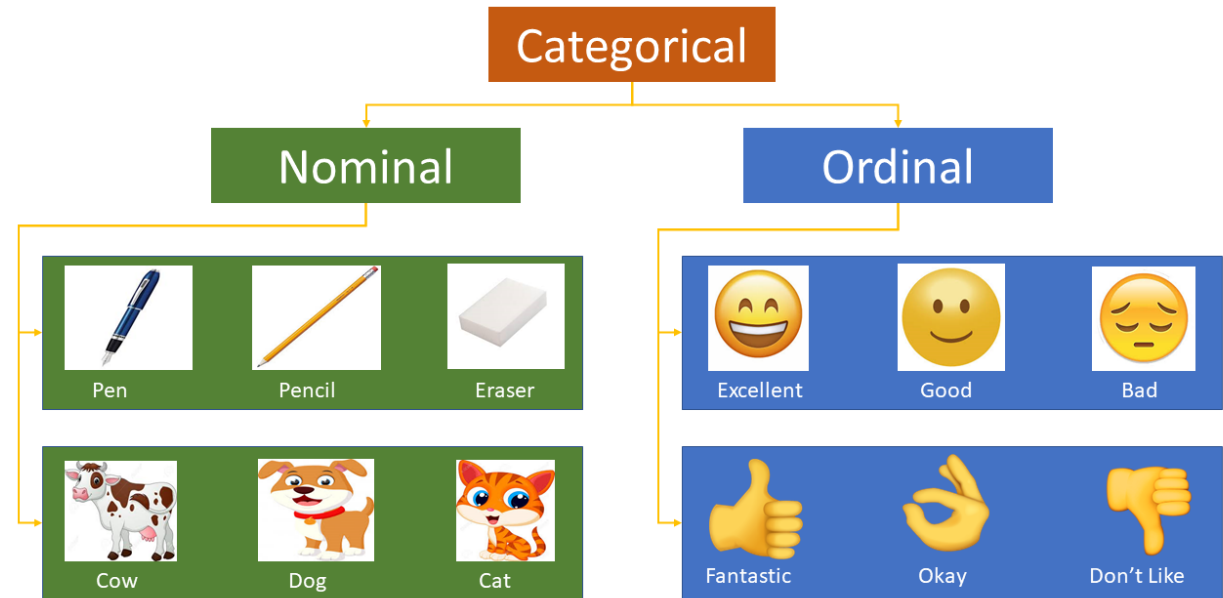
O que são fatores?

Variáveis qualitativas

- ▶ Nominais ou ordinais.
- ▶ São propriedades ou atributos que expressam qualidade.
- ▶ Nominais: a ordem não importa.
- ▶ Ordinais: existe ordenação natural.

Exemplos:

- ▶ Nota para avaliação de produto/serviço.
- ▶ Escala likert.
- ▶ Unidades geográficas como UF.
- ▶ Unidades cronológicas como meses.
- ▶ Tipo de produto/departamento.



Tipos de variáveis qualitativas. Fonte:
<https://morioh.com/p/c7649e4e463f>.

Conteúdo do pacote

- ▶ Página do pacote: <https://forcats.tidyverse.org/>.
- ▶ Cheatsheet: <https://github.com/rstudio/cheatsheets/blob/main/factors.pdf>

Principais funções

- ▶ `fct_reorder()`: **reordena** os níveis baseado em outra variável. Ex: reordenar UFs pelo número de habitantes.
- ▶ `fct_relevel()`: troca os níveis do fator de **posição**. Ex: passar o nível "desconhecido" para última posição.
- ▶ `fct_relabel()`: para **rerotular** os níveis. Ex: São Paulo → SP, Mato Grosso do Sul → MS.
- ▶ `fct_recode()`: para **recodificar** os níveis do fator. Ex: {PR, SC, RS} → Sul, {MS, MT, GO} → Centro-oeste.
- ▶ `fct_lump*()`: **aglutinar** níveis conforme critérios. Ex: agrupar os níveis de menor ocorrência sob o rótulo "outros".

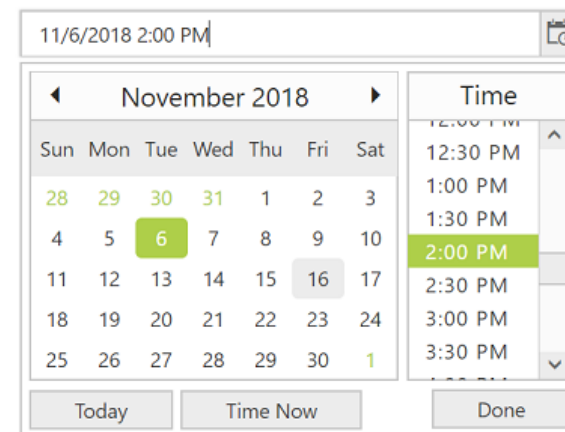
A maioria das funções do pacote tem o prefixo `fct_`.

<code>%>%</code>	<code>fct_other</code>
<code>as_factor</code>	<code>fct_recode</code>
<code>fct_anon</code>	<code>fct_relabel</code>
<code>fct_c</code>	<code>fct_relevel</code>
<code>fct_collapse</code>	<code>fct_reorder</code>
<code>fct_count</code>	<code>fct_reorder2</code>
<code>fct_cross</code>	<code>fct_rev</code>
<code>fct_drop</code>	<code>fct_shift</code>
<code>fct_expand</code>	<code>fct_shuffle</code>
<code>fct_explicit_na</code>	<code>fct_unify</code>
<code>fct_infreq</code>	<code>fct_unique</code>
<code>fct_inorder</code>	<code>first2</code>
<code>fct_inseq</code>	<code>gss_cat</code>
<code>fct_lump</code>	<code>last2</code>
<code>fct_lump_lowfreq</code>	<code>lvls_expand</code>
<code>fct_lump_min</code>	<code>lvls_reorder</code>
<code>fct_lump_n</code>	<code>lvls_revalue</code>
<code>fct_lump_prop</code>	<code>lvls_union</code>
<code>fct_match</code>	

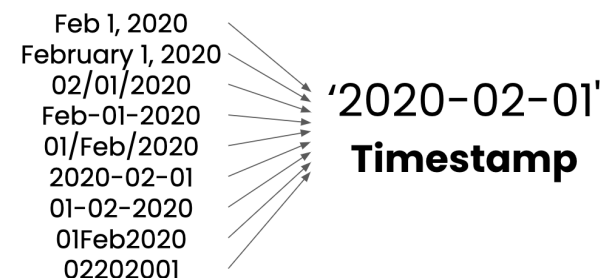
{lubridate}

O que são variáveis de data e tempo?

- ▶ Datas e tempos registram o instante, duração, período de eventos.
- ▶ São complicados por serem irregulares:
 - ▶ Convenções de representação dependem do idioma.
 - ▶ Existem os fusos horários e horários de verão.
 - ▶ Anos podem ser bissextos.
 - ▶ Existem feriados que mudam com os países e estados.
 - ▶ Meses de tamanhos diferentes.
 - ▶ Horas e minutos não usam base decimal.
- ▶ Recursos consistentes e acessíveis para trabalhar com datas e tempos são muito bem-vindos.
- ▶ Página do pacote: <https://lubridate.tidyverse.org/>.
- ▶ Cheatsheet: <https://github.com/rstudio/cheatsheets/blob/main/lubridate.pdf>



Seletor de data e hora.



Tipos de representação de datas. Fonte:
<https://www.dataindependent.com/pandas/pandas-to-datetime/>.

Conteúdo do pacote

%--%	dmilliseconds	hours	leap_year	period	tz<-
%m-%	dminutes	int_aligns	local_time	period_to_seconds	union
%m+%	dmonths	int_diff	make_date	picoseconds	wday
%within%	dmy	int_end	make_datetime	pm	wday<-
add_with_rollback	dmy_h	int_end<-	make_difftime	POSIXct	week
am	dmy_hm	int_flip	mday	pretty_dates	week<-
Arith	dmy_hms	int_length	mday<-	qday	weeks
as_date	dnanoseconds	int_overlaps	mdy	qday<-	with_tz
as_datetime	dpicoseconds	int_shift	mdy_h	quarter	yday
as.difftime	dseconds	int_standardize	mdy_hm	reclass_date	yday<-
as.duration	dst	int_start	mdy_hms	reclass_timespan	ydm
as.interval	duration	int_start<-	microseconds	rollback	ydm_h
as.period	dweeks	intersect	milliseconds	rollbackward	ydm_hm
ceiling_date	dyears	interval	minute	rollforward	ydm_hms
Compare	dym	is.Date	minute<-	round_date	year
cyclic_encoding	epiweek	is.difftime	minutes	second	year<-
date	epiyear	is.duration	month	second<-	years
Date	fast_strptime	is.instant	month<-	seconds	ym
date_decimal	fit_to_timeline	is.interval	ms	seconds_to_period	ymd
date<-	floor_date	is.period	my	semester	ymd_h
day	force_tz	is.POSIXct	myd	setdiff	ymd_hm
day<-	force_tzs	is.POSIXlt	NA_Date_	show	ymd_hms
days	format_ISO8601	is.POSIXt	NA_POSIXct_	stamp	yq
days_in_month	guess_formats	is.timepoint	nanoseconds	stamp_date	
ddays	hm	is.timespan	now	stamp_time	
decimal_date	hms	isoweek	origin	time_length	
dhours	hour	isoyear	parse_date_time	today	
dmicroseconds	hour<-	lakers	parse_date_time2	tz	

Condições finais

- ▶ Dados como datas e fatores são muito frequentes, bem como a necessidade de tratar cadeias de caracteres.
- ▶ O {stringr} oferece recursos para emprego de regex na manipulação de campos de texto.
- ▶ O {forcats} reúne várias funções de conveniência para trabalhar com fatores.
- ▶ O {lubridate} traz funcionalidades para trabalhar variáveis de datas.

Opinião do Dono

dados semi-estruturados

"Melhor hatch pequeno"
Volkswagen Gol G7 Track 1.0 2017/2017 ← *categórica*
Renato - Campo Grande MS ← *geográfica*
Dono há 1 ano - 11.090 km ← *duração*
Carro anterior: Ford Ecosport ← *transação*

dados não estruturados

Prós:
Aparência externa muito bonita, moderna, interior show, igual ao do Golf, painel completo, bonito, com computador de bordo, multimídia, rodas de liga leve, controles no volante, econômico, desempenho surpreendente, nem parece 1.0. A grade dianteira dele, aliada aos faróis de design moderno, dão um ar mais agressivo e esportivo ao carro, deixando ele muito atraente.

Contras:
Porta malas pequeno, espaço para quem viaja atrás, no meio, bastante desconfortável, ruídos nas portas, alguns ruídos o painel, suspensão dura.

Defeitos apresentados:
Até agora, somente caiu o chip da chave canivete, que impossibilitava o funcionamento do carro.

Opinião Geral:
A aparência desse Gol me surpreendeu positivamente, um carro muito bonito, sobretudo o interior. Quem estava acostumado com o Gol G5, vai se surpreender também. O motor de 3 cilindros, a princípio me deixou um tanto preocupado, mas o desempenho é excelente, bastante ágil, além de deixa-lo muito econômico. O preço é salgado, paguei R\$ 45.000,00, contudo, a revenda é garantida, pois o Gol ainda é um dos carros usados mais procurados no mercado. Boa compra, eu recomendo.

11/02/2018 23:25:00 ← *cronológica*

dados estruturados

Estilo	★★★★★	10
Acabamento	★★★★☆	7
Posição de dirigir	★★★★☆	7
Instrumentos	★★★★★	10
Interior	★★★☆☆	6
Porta-malas	★★☆☆☆	5
Desempenho	★★★★★	9
Motor	★★★★★	8
Câmbio	★★★★★	9
Freios	★★★★★	8
Suspensão	★★★☆☆	6
Consumo	★★★★★	10
Estabilidade	★★★★★	8
Custo-Benefício	★★★★★	8
Recomendação	★★★★★	8
Avaliação Geral	★★★★★	7,93

discreta limitada

Avaliação online de um veículo do site carros na web.

Fonte: <https://github.com/leg-ufpr/hackathon/blob/master/datasets.md>.