

Ciencia de datos, práctica 6

Juan Casado Ballesteros, Samuel García Gonzalez, Iván Anaya Martín

December 10, 2019

Abstract

Analizaremos las gráficas que permiten visualizar la regresión. En ellas deberemos mostrar los datos como una nube de puntos junto a la recta de regresión que los aproxima.

Realizaremos una pequeña guía de estilo incluyendo consejos sobre como crear una buena gráfica en general y en concreto una de regresión. Para hacerlo nos hemos inspirado en el libro de Edward Tufte, *Data Analysis for Politics and Policy* donde habla en el capítulo 3 sobre las rectas de regresión.

Utilizaremos algunas técnicas de las que él utiliza para mejorar la visualización de la regresión como la esca logarítmica. También expondremos algunos de sus más famosos ejemplos como apoyo a explicar la importancia de la visualización de la regresión.

Finalmente concluiremos con ejemplos de rectas de regresión dibujadas por nosotros con las funciones de R siguiendo los principios que nosotros mismos hemos propuesto. También realizaremos ejemplos de rectas de regresión con los paquetes "plotly" y "ggplot2". Estos paquetes tienen la capacidad de crear gráficas interactivas lo cual también comentaremos y ejemplificaremos mencionando sus ventajas frente a las gráficas estáticas.

Contents

1	Visualización de la regresión	2
1.1	Ejemplos de regresión	2
1.2	Mejora de visualización en gráficas con recta de regresión	3
2	La importancia de la visualización	4
2.1	Cantidad de información mostrada	5
2.2	Paquetes de visualización	8
2.2.1	Representaciones interactivas	10

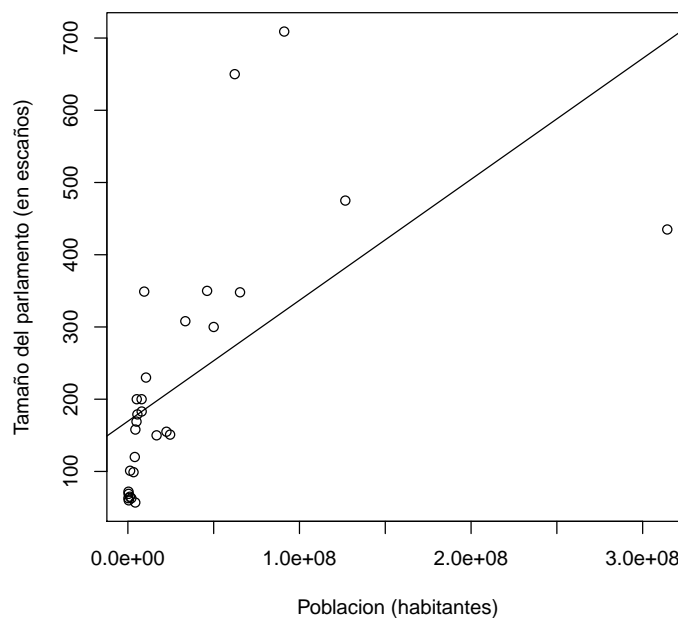
1 Visualización de la regresión

La recta de ajuste entre dos variables es una herramienta muy importante en el análisis de datos y su visualización, aunque no por si sola, si la correlación entre las variables es baja esta puede no significar nada. Lo realmente relevante son la relación entre la recta y los datos. La visualización es importante pues nos permite condensar la información y simplificarla para permitir entenderla de forma rápida. También es interesante porque podemos extraer conclusiones que a primera vista no hemos sido capaces de observar numéricamente. Deben poder visualizarse ambos elementos, la recta y la nube de puntos de los datos de forma adecuada pues es la relación entre ambos la que nos interesa, deseamos ver lo bien o mal que la recta se adapte a nuestros datos.

1.1 Ejemplos de regresión

En el análisis de regresión compararemos como dos variables de la siguiente muestra se relacionan entre ellas. Estos datos representan la relación existente entre la población de un país y el tamaño de su parlamento. Hemos escogido datos de 29 países en los que su sistema político se basa en la democracia liberal.

Hay que tener en cuenta que una buena representación de datos debe ser limpia y que a la vez facilite la comprensión de la mayor cantidad de información en el menor tiempo posible. Este objetivo lo queremos cumplir en todas las representaciones de este apartado.

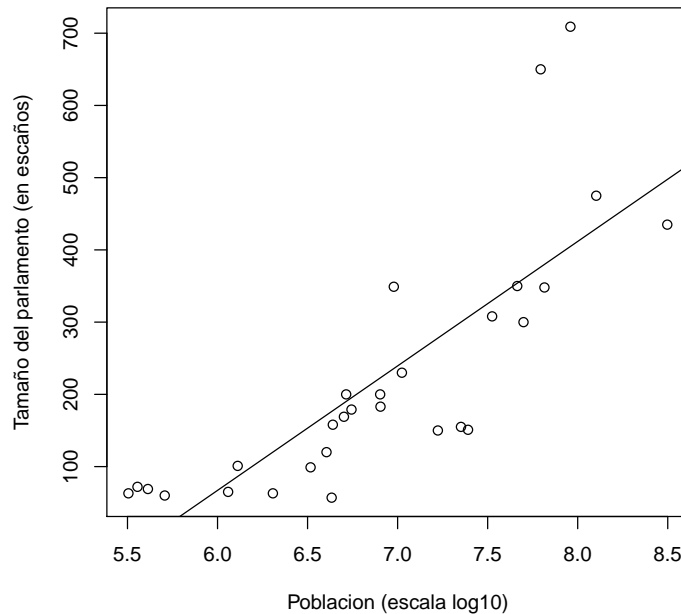


1.2 Mejora de visualización en gráficas con recta de regresión

Es muy importante indicar las unidades de medida en las que representamos los datos, porque modificando estas (sin manipular los datos) podemos conseguir una mejor o peor representación de los mismos. Una manera de mejorar esta gráfica es representando las variables de manera logarítmica. La usamos con los siguientes propósitos, interpretados de las conclusiones de Edward Tufte en su libro *Data Analysis for Politics and Policy*:

- Los coeficientes de regresión a veces son más útiles en una interpretación teórica.
- Cuando una gráfica está mal distribuida a primera vista y sus datos están agrupados en diferentes clusters, junto con algún dato periférico, se pueden transformar usando en la representación el logaritmo de las observaciones consiguiendo de esta manera que los valores de los clusters se “extiendan” en la gráfica y los valores periféricos se “concentren” también, obteniendo una distribución mucho mejor.
- Algunos supuestos teóricos que son fundamentales en el modelo de regresión y su significado en los resultados encajan mejor cuando se usa el logaritmo de las unidades de medida.

Estos aspectos que enuncia Tufte, nos benefician en la representación de la muestra que tenemos, puesto que según esas conclusiones (las cuales vamos a probar que son ciertas), extraemos que la representación logarítmica de las variables da resultado a una distribución mucho más simétrica, puesto que los valores muy grandes los “aplana” hacia la media, y los valores pequeños los “estira”. Cabe destacar que transformar las variables usando logaritmos reduce la no-linealidad en la relación y reduce en gran medida el desorden en las distribuciones, es decir, ayuda a clarificar la relación entre dos variables. Pasamos a probar estos fundamentos teóricos en la realidad, representando la población en escala logarítmica:



Como podemos ver, es mucho más clara no solo la relación entre las dos variables, pues en la anterior gráfica podíamos apreciar que los datos estaban relacionados, sino también la distribución de puntos, y con mucha diferencia, puesto que en la anterior gráfica se concentraban mucho en la parte inferior. En la nueva podemos ver que están mucho más repartidos y la comprensión a primera vista es más clara y rápida. Este análisis nos demuestra que también se puede mejorar la visualización de una gráfica usando otros métodos aparte de la mejora simplemente visual, como hemos hecho en este apartado modificando los ejes de representación.

2 La importancia de la visualización

Como ejemplo ilustrativo de la importancia de la visualización ponemos el ya famoso ejemplo del libro de Edward Tufte, *Data Analysis for Politics and Policy*, Chapter 3: Two-Variable Linear Regression. En dicho ejemplo las rectas de regresión obtenidas son idénticas y se adaptan a ojos de la correlación de igual modo a los datos. Si no fuera por la visualización de ambos elementos en conjunto no podríamos conocer la gran disparidad entre cada una de las muestras.

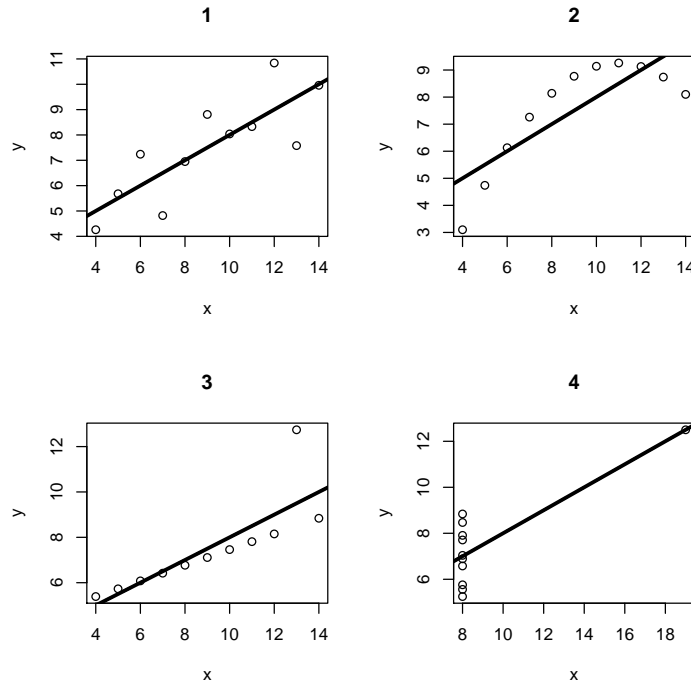
Si representáremos los datos solos sería difícil imaginar que tuvieran la misma regresión y si visualizáramos las regresiones solo sería difícil imaginar que podrían estar representando a datos tan dispares.

[1] "1: Correlacion cuadrada: 0.6665425 a: 3.000091 b: 0.5000909"

[1] "2: Correlacion cuadrada: 0.666242 a: 3.000909 b: 0.50"

[1] "3: Correlacion cuadrada: 0.666324 a: 3.002455 b: 0.4997273"

[1] "4: Correlacion cuadrada: 0.6667073 a: 3.001727 b: 0.4999091"



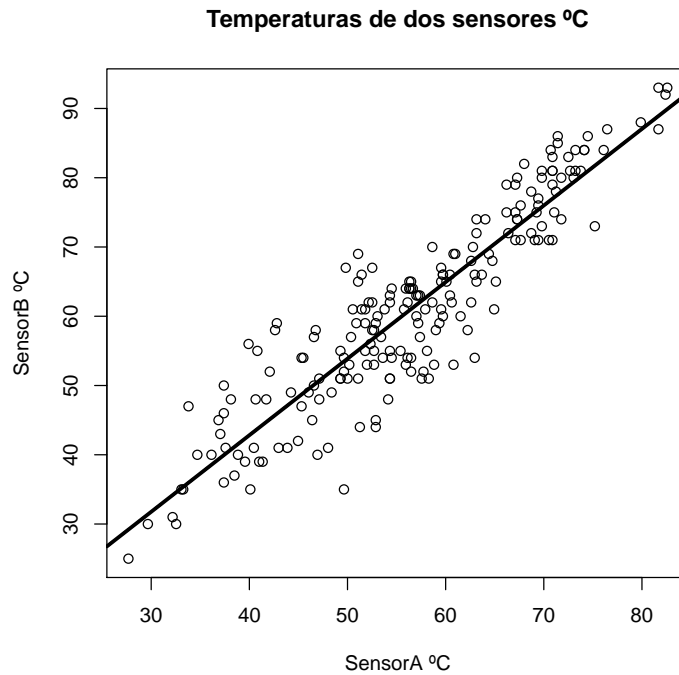
2.1 Cantidad de información mostrada

Debido a la simplicidad de esta gráfica, solo se está mostrando una nube de puntos y la recta de regresión es tentador comenzar a añadir elementos adicionales. No obstante antes de hacerlo debemos tener en cuenta las siguientes consideraciones.

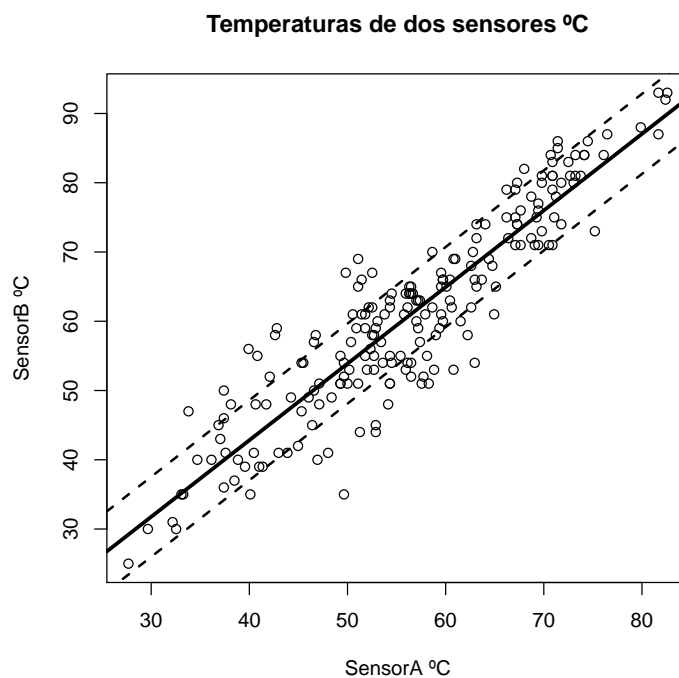
- Se debe mostrar la información mínima necesaria para mostrar aquello que deseamos. Si tenemos que mostrar una relación compleja entre los datos puede que sea necesario utilizar una representación compleja. No obstante si podemos hacerla simple será mejor pues podrá ser entendida por más gente de forma más rápida.
- Debemos de indicar el significado de cada elemento que añadamos a la gráfica cuando este no sea claro.
- Es de gran importancia indicar la magnitud de los datos tanto como el dato que se está representando en cada eje.

[1] "Correlacion cuadrada de Temperaturas: 0.835327071582904"

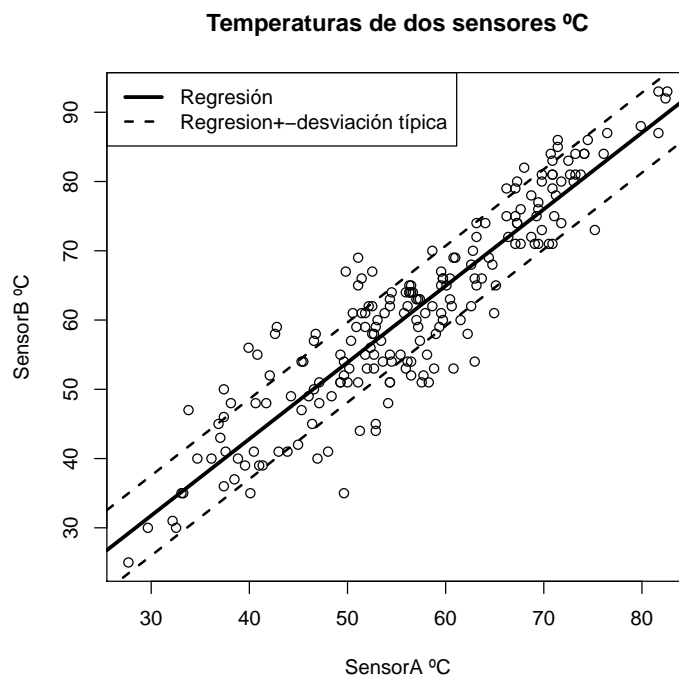
En este caso mostramos exclusivamente los datos junto a su recta de regresión. Obtenemos una representación muy clara y sencilla que nos muestra directamente la información que queríamos. La gráfica tiene un título y en cada eje mostramos la variable representada y las unidades de esta.



Podemos aumentar la complejidad de la representación añadiendo dos líneas paralelas a la recta de regresión que nos indiquen la desviación típica de esta. Este elemento adicional nos proporciona información adicional sobre la regresión que dependiendo del contexto puede ser necesaria si a partir de los datos y de la recta es difícil juzgar la calidad de esta. No obstante podemos ver como solo con haber añadido un elemento tan simple la representación se siente mucho más densa.



Como hemos indicado para este caso sería recomendable añadir una leyenda que nos indicara qué datos se están representando de modo que la visualización obtenida sea más clara.



2.2 Paquetes de visualización

Para realizar las gráficas anteriores hemos utilizado la siguiente función que utiliza las funciones propias de R.

```
> regPlot

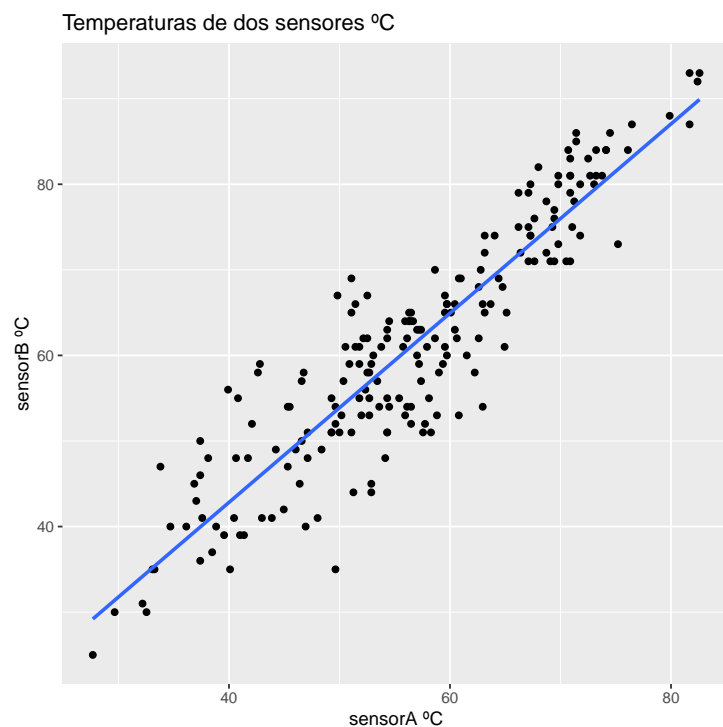
function (x, y, regresion, limit, title="", xlabel="", ylabel="") {
  plot(x, y, xlab=xlabel, ylab=ylabel, main=title)
  regUpLimit <- regresion
  regUpLimit$coefficients[1] = regUpLimit$coefficients[1] + limit
  regDownLimit <- regresion
  regDownLimit$coefficients[1] = regDownLimit$coefficients[1] - limit
  abline(regUpLimit, "gray", lty=2, lwd=2)
  abline(regresion, "black", lty=1, lwd=3)
  abline(regDownLimit, "gray", lty=2, lwd=2)
}
<bytecode: 0x7fd63f7412c8>
```

Mediante plot visualizamos la nube de puntos y con abline dibujamos líneas rectas sobre la última gráfica creada.

Existen una gran cantidad de paquetes que se pueden utilizar para visualizar datos. Uno de los más famosos es "ggPlot2". Esto se debe a que produce gráficas muy vistosas, altamente configurables todo con una sintaxis sencilla y fácil de aprender.

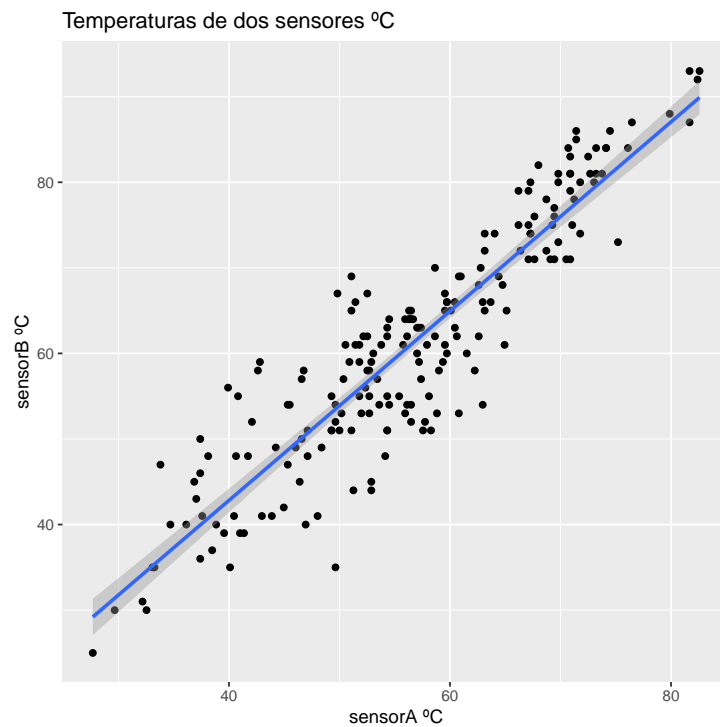
No obstante esto también es uno de los problemas de esta librería. Por defecto las gráficas que obtendremos tendrán un fondo grisáceo con una cuadrícula en blanco tal y como podemos ver en el ejemplo. Esto hace que la representación de los datos sea más difusa que la obtenida con las funciones propias de R que crean una visualización mucho más limpia.

```
> ggplot(data = datos3, aes(x = datos3$Temperature_ElMonte,
+                             y = datos3$Temperature_Sandburg)) +
+   geom_point(color='black') +
+   geom_smooth(method = "lm", se = FALSE) +
+   labs(title = "Temperaturas de dos sensores °C") +
+   xlab("sensorA °C") + ylab("sensorB °C")
```

Esta librería permite además añadir gran variedad de elementos de forma sencilla a las gráficas. Dichos elementos se ven siempre muy vistosos y bonitos pero pueden no tener gran significado para nuestro caso concreto. Debemos por tanto tener cuidado y no abusar de ellos.

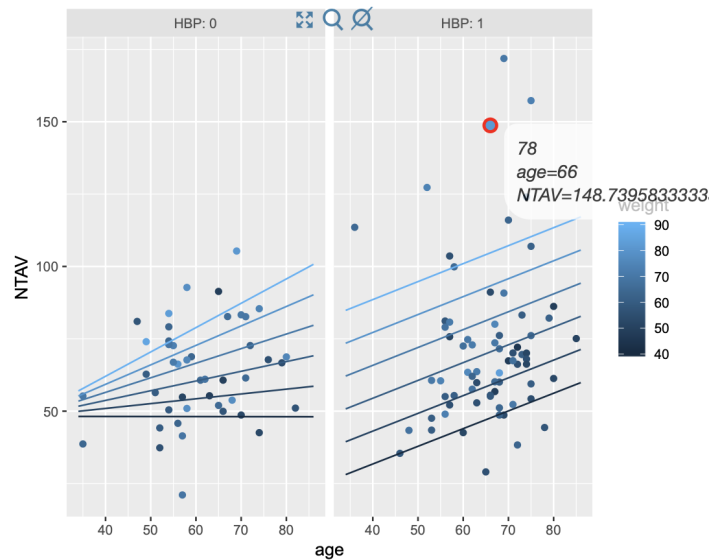
```
> ggplot(data = datos3, aes(x = datos3$Temperature_ElMonte,
+                             y = datos3$Temperature_Sandburg)) +
+   geom_point(color='black')+
+   geom_smooth(method="lm")+
+   labs(title = "Temperaturas de dos sensores °C") +
+   xlab("sensorA °C") + ylab("sensorB °C")
```



2.2.1 Representaciones interactivas

Una de las ventajas que tenemos con estas librerías es que podemos crear gráficas interactivas. Estas gráficas crean un archivo .html que podemos visualizar de forma interactiva en el navegador. Las gráficas interactivas nos permiten realizar representaciones más complejas de los datos así como representar más datos a la vez. Esto se debe a que cuando pasemos el cursor encima de los elementos se resaltarán los que están relacionados pudiendo ver simultáneamente estos con respecto a los otros.

Mostramos una gráfica creada con "ggplot2" que se ha convertido en una gráfica interactiva con "ggiraphExtra".



En este caso mostramos una gráfica interactiva creada con "plotly" la cual utiliza "ggplot2" por debajo para crear sus gráficas. Esta librería tiene también una sintaxis sencilla y permite crear gráficas muy configurables. Como ventaja frente a "ggplot2" debemos comentar que los estilos que se aplican por defecto a las gráficas son mucho más limpios y claros.

```
> datos3 %>%
+   plot_ly(x = ~Temperature_ElMonte, y = ~Temperature_Sandburg) %>%
+   add_markers (x = ~Temperature_ElMonte, y = ~Temperature_Sandburg,
+   name="temperaturas °C") %>%
+   add_lines(x = ~Temperature_ElMonte, y = fitted(regresion3), name="regresion")%>%
+   layout(title = "Temperaturas de dos sensores °C",
+   xaxis = list(title = "sensorA °C"),
+   yaxis = list(title = "sensorB °C"))
```

