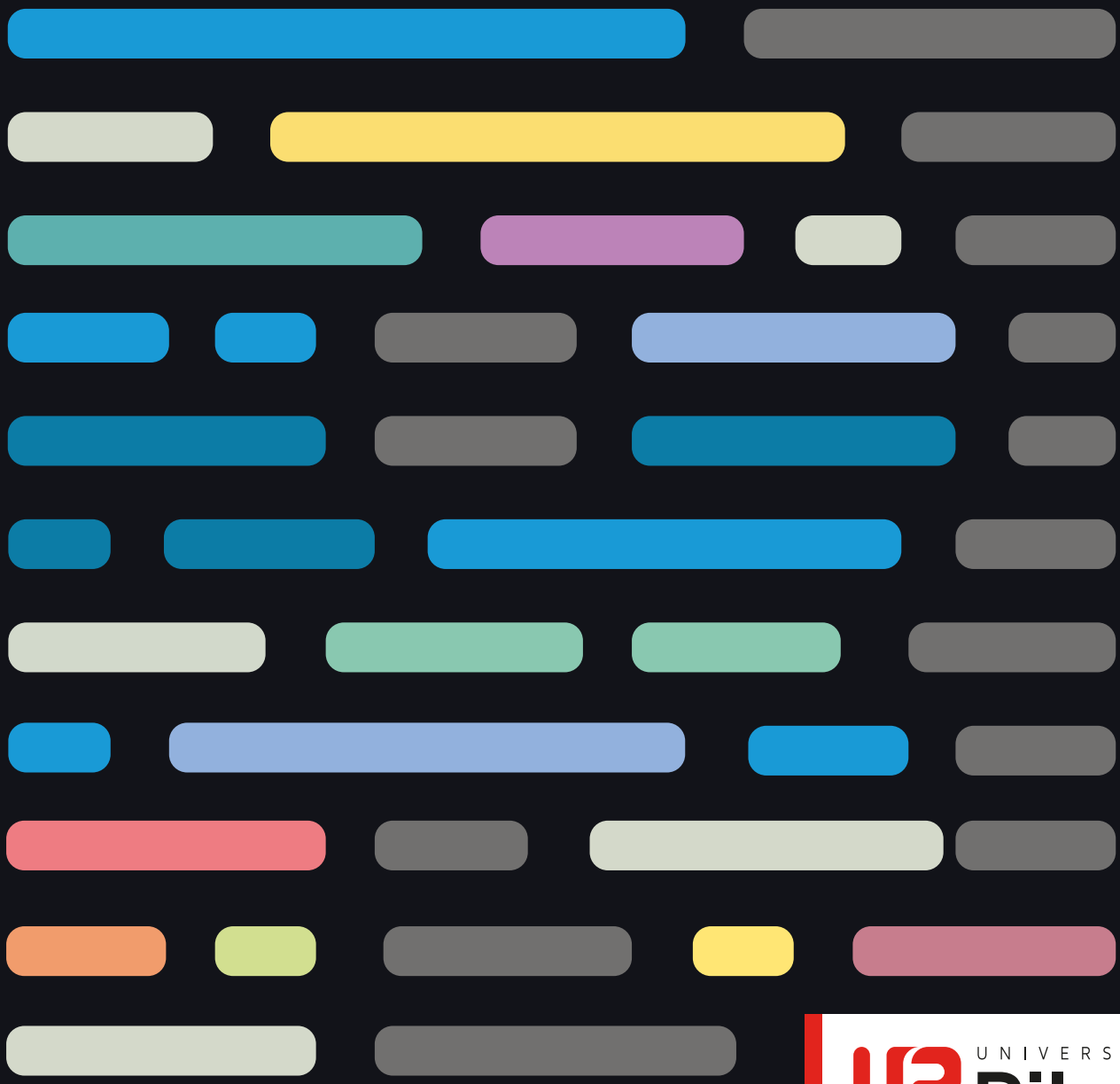
Introducción al análisis de datos usando el lenguaje de programación R

Margarita Zuluaga



UNIVERSIDAD Piloto DE COLOMBIA

Zuluaga, Margarita

Introducción al análisis de datos usando el lenguaje de programación R / Margarita Zuluaga. Bogotá: Universidad Piloto de Colombia, 2023

69 Páginas.

Incluye referencias.

ISBN: 9789585106949

Estadística – Procesamiento de datos

Estadística – Programas para computador

R (Lenguaje de programación para computadores)

Análisis de datos estadísticos – Lenguajes de programación

I. Rondón Quintana, Hugo Alexander autor

II. Bastidas Martínez, Juan Gabriel autor

CCD: 005.13

Autor
Margarita Zuluaga

ISBN
978-958-5106-94-9

Primera edición – 2023
Bogotá, Colombia

Coordinadora editorial
Catalina Moreno Correa

Diseño
María Paula Martín Peña

Contenido

Introducción

R es un lenguaje de programación para análisis estadístico y gráfico. Este lenguaje de programación es de uso libre y contiene una gran cantidad de librerías que permiten realizar tratamientos estadísticos en modelos lineales y no lineales, test estadísticos, análisis de series temporales, algoritmos de clasificación, agrupamientos y permite la visualización de los datos en gráficas como histogramas, diagramas de tallo y hojas, gráfico de cajas y bigotes, de simetría, de dispersión o en mapas, entre otras. El avance en la tecnología ha hecho notorio el incremento en la producción y análisis computacional de grandes volúmenes de datos trayendo transformaciones en la forma en la cual se pueden tomar decisiones. El mundo requiere personas entrenadas que puedan analizar datos y contribuir en el proceso de toma de decisión a nivel gerencial, empresas privadas o en políticas públicas. Para los futuros egresados de la Universidad Piloto de Colombia (UniPiloto) es un complemento a su formación, ya que les puede permitir integrar sus conocimientos a través de proyectos pequeños en R donde se plasmen conceptos aprendidos tanto en el componente básico como el componente de profundización y electivas.

Existen cuatro habilidades que entran en acción en cada proyecto que involucra ciencia de datos, en mayor o menor medida, de acuerdo con la etapa de análisis: programación, estadística, comunicación y conocimiento de dominio (Vazquez, 2019). Al hacer análisis de datos se ponen en práctica algunas técnicas de programación para resolver problemas. La estadística se hace necesaria para extraer conocimiento de los datos (media, mediana, desvío estándar y cuartiles). Con la comunicación se busca explicar las revelaciones de un modelo estadístico a términos que tengan sentido para un público amplio, crear visualizaciones que permitan a terceros "leer" los datos y sacar conclusiones por su cuenta. El conocimiento de dominio es la experiencia acumulada en un campo particular de actividad humana, este ayuda a discernir si las respuestas obtenidas mediante un sofisticado análisis estadístico tienen sentido. Para los estudiantes que van a presentar su prueba Saber Pro es una oportunidad de sintetizar conceptos aprendidos a lo largo de su proceso académico en el pregrado.

El material está pensado para los estudiantes de la Universidad Piloto y el público en general que quiera aprender a programar con R y realizar ciencia de datos para proyectos personales o para presentar a empresas del sector público o privado. Muchas entidades ponen a disposición en sus portales datos que se pueden transformar para visualizarlos de una manera más atractiva y así ayudar a las mismas empresas en la toma de decisiones. Los estudiantes de la UniPiloto que están próximos a presentar sus pruebas Saber Pro pueden aprovechar este material como una herramienta adicional para relacionar conceptos aprendidos a lo largo de su formación profesional usando una aplicación práctica de manejo de datos a través de la programación con R.

Conociendo R

R es un lenguaje de programación usado en la ciencia de datos. Es un producto de código abierto que permite un libre acceso a la herramienta sin necesidad de pagar licencia. Al ser de uso libre es posible realizar colaboraciones en proyectos ya iniciados o alimentar la herramienta diseñando o programando "paquetes" para realizar algunas tareas específicas dentro de R.

R ofrece un manejo y almacenamiento efectivo de los datos, permite realizar cálculos con matrices, librerías para el análisis de datos, herramientas gráficas para la visualización de datos, un formato para visualizar los datos en diferentes formatos como PDF, Word o en línea. Permite la manipulación, procesamiento y visualización gráfica de los datos creando visualizaciones de alta calidad, dashboards para visualización y análisis de datos, creación de informes y análisis estadísticos para ahondar en el conocimiento de los datos (Lenguaje R, ¿qué es y por qué es tan usado en big data?, 2019).

La ciencia de datos ofrece la oportunidad para tratar los datos que en sí no aportan información relevante a menos que sean analizados para convertirlos en conocimiento y con ello aportar criterios a empresas y organizaciones en la toma de decisiones (Vázquez, 2019).

Dentro del lenguaje de R se ejecutan varias tareas, la primera de ellas es la adquisición de datos que se pueden encontrar en diferentes fuentes de acceso abierto al público como portales gubernamentales, bases de datos, redes sociales, entre otras, para darles una estructura adecuada. Luego se hace la preparación de estos, se realiza una exploración aplicando técnicas como correlación, tendencias, entre otras, para entender los datos mediante un análisis preliminar. Dentro de esta preparación también se hace una limpieza de valores incoherentes, duplicados, inválidos, se agrupan en estructuras útiles o manejables para su procesamiento.

Ya con los datos organizados viene la fase de análisis mediante la selección de las técnicas adecuadas y la construcción de modelos predictivos, clasificación, agrupación, entre otras. Una vez cumplida esta etapa viene la comunicación de los resultados y las aplicaciones del modelo desarrollado (Analista de datos: ¿cuál es el perfil de estos profesionales?, 2019).

R es un lenguaje de programación ideal para empezar a desarrollar habilidades en el campo de la ciencia de datos. Además, ofrece un ambiente interactivo mucho más flexible que otros lenguajes de programación que permite desarrollar una gramática que ayuda a pensar e interactuar de manera fluida con la programación en R (Wickham y Grolemund, 2019).

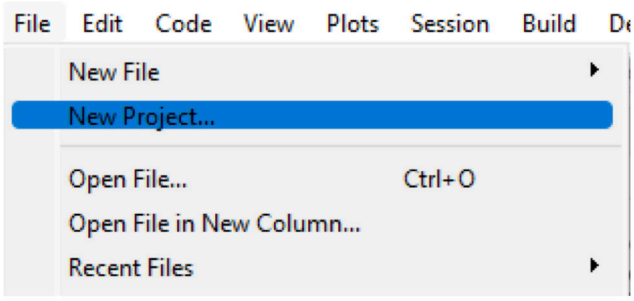
Primer proyecto en R

La ciencia de datos permite, por ejemplo, tratar datos urbanos para tomar mejores decisiones a la hora de desarrollar políticas públicas. Los datos se pueden obtener de manera libre a través de telefonía celular, apps móviles, redes sociales, entre otras. La información recolectada luego se puede organizar de tal forma que aporte valor al relacionar datos, encontrar patrones, entender comportamientos y con esto aportar a la toma de decisiones en políticas públicas mejorando la calidad de vida de las personas.

Para iniciar en el computador se instala el programa R, esto se hace entrando a su web (http://www.r-project.org) y siguiendo las instrucciones. También puede ser útil usar un video tutorial como el que se encuentra en https://www.youtube.com/watch?v=R-7cDi6BNGk&list=PLdV8ntSOIL5SqS4-sbms4M_pullucPwmh&index=4

Después de tener instalado el programa R, el primer paso es ejecutar Rstudio; en la interfaz gráfica se crea un nuevo proyecto dando clic en

File -> New Project... -> New Directory -> New Project.



► **Figura 1.** Interfaz gráfica para crear un nuevo proyecto en R

En la ventana que surge, elegir un nombre para el proyecto (por ejemplo, "Practicando R") y finalizar la operación clicando en "Create project".

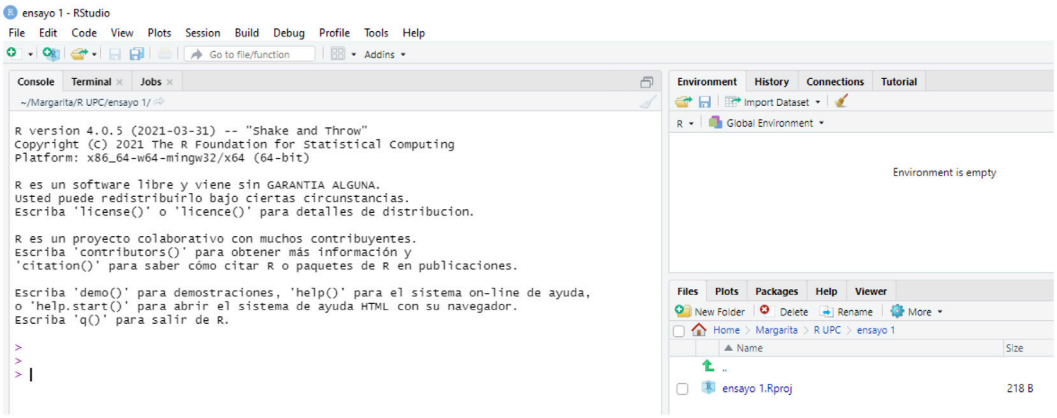
Utilizar proyectos permite continuar en otro momento desde donde quedó la tarea al terminar una sesión. Al abrir RStudio, clicando en

File -> Recent Projects -> "nombre de mi proyecto".

A continuación, se debe crear un script. Un script, como su nombre en inglés lo indica, es un guion; una serie de pasos que escribimos para que nuestra computadora ejecute en secuencia.

Clicar en File -> New File -> R Script.

De inmediato se abre una ventana con un editor de texto.



► **Figura 2.** Creación de un script en R

Para practicar, se plantean los siguientes ejercicios donde se usan algunas de las librerías en R que permiten el análisis y la visualización de los datos. En los ejercicios de visualización de datos presentados a continuación se van a usar las siguientes librerías:

La primera librería que se va a cargar es tidyverse, esta librería es un conjunto de paquetes entre los cuales se encuentran ggplot2, para visualización de datos; dplyr, para la manipulación de datos; tidyr, para ordenar los datos; readr, para la importación de datos, y stringr, para cadenas.

Al cargar la librería tidyverse aparece también el listado de los paquetes que pueden estar en conflicto entre paquetes de tidyverse y otros paquetes que se hayan cargado. Los que aparecen aquí en este ejemplo se ignoran ya que los equivalentes básicos son genéricos (Wickham, s. f.).

La siguiente librería es: library(sf)

Con esta librería se consigue un paquete que proporciona acceso a funciones simples para R que permiten describir los objetos del mundo real haciendo un énfasis en la geometría espacial de estos objetos registrando en el data.frame datos simples con una columna de lista geométrica, dando acceso a trabajar con bases de datos espaciales como GeoJSON.

Para hacer una inspección preliminar de un data.frame o un objeto se puede usar summary ().

Una vez se tenga la información organizada se puede usar ggplot para graficar los datos. Los elementos de ggplot se unen con el símbolo + y se pueden agregar los elementos básicos de ggplot como geom_point(): indica la geometría o tipo de gráfico.

La librería lubridate () permite trabajar con datos que involucran fechas y horas. Algunas funciones simples para establecer componentes de una fecha y hora son year (), month (), mday (), hour () y: minute () second ().

Ejercicio 1. Proyecto en R: Bogotá

Emergencias reportadas en Bogotá entre enero de 2018 y febrero de 2021

Muchas ciudades tienen portales que permiten consultar datos abiertos de la ciudad. A continuación, una lista con algunas ciudades y el enlace:

ID	Ciudad	Pais	Link
1	Berlín	Alemania	https://daten.berlin.de/
2	Buenos Aires	Argentina	https://data.buenosaires.gob.ar
3	Córdoba	Argentina	https://gobiernoabierto.cordoba.gob.ar/data/datos-abiertos
4	Rosario	Argentina	https://datos.rosario.gob.ar/
5	Mar del Plata	Argentina	https://datos.mardelplata.gob.ar
6	Bahía Blanca	Argentina	https://datos.bahia.gob.ar/
7	Salta	Argentina	http://geoportal.idesa.gob.ar/
8	Toronto	Canadá	https://open.toronto.ca/
9	Barcelona	España	https://opendata-ajuntament.barcelona.cat/es/
10	Madrid	España	https://datos.comunidad.madrid/catalogo/
11	Bilbao	España	https://www.bilbao.eus/opendata/es/catalogo
12	París	Francia	https://opendata.paris.fr/pages/home/
13	Ámsterdam	Holanda	https://data.amsterdam.nl/
14	Singapur	Singapur	https://data.gov.sg
15	Zúrich	Suiza	https://data.stadt-zuerich.ch/
16	Montevideo	Uruguay	https://catalogodatos.gub.uy/organization/intendencia-montevideo
17	Boston	EE. UU.	https://data.boston.gov/dataset

18	New York	EE. UU.	https://opendata.cityofnewyork.us
19	San Francisco	EE. UU.	https://datasf.org/opendata/
20	Chicago	EE. UU.	https://data.cityofchicago.org/
21	Detroit	EE. UU.	https://data.detroitmi.gov/
22	Sydney	Australia	https://data.cityofsydney.nsw.gov.au/
23	Londres	Inglaterra	https://data.london.gov.uk/
24	Ciudad de México	México	https://datos.cdmx.gob.mx/
25	São Paulo	Brasil	http://dados.prefeitura.sp.gov.br/
26	Medellín	Colombia	https://geomedellin-m-medellin.opendata.arcgis.com/
27	Mendoza	Argentina	https://datos.ciudaddemendoza.gob.ar/
28	Cape Town (Ciudad del Cabo)	Sudáfrica	https://odp-cctegis.opendata.arcgis.com/
29	Múltiples	Dinamarca	https://www.opendata.dk/

En la página de la Alcaldía de Bogotá se encuentra un portal con datos abiertos. Para este ejercicio se ingresa a la página:

https://datosabiertos.bogota.gov.co/dataset/bitacora-de-emergencias

De acuerdo con el problema que se quiera resolver, se seleccionan los archivos cvs que pueden contener información relevante.

Para este ejercicio se consultó el reporte de emergencias en la ciudad de Bogotá en un periodo comprendido entre 01 de enero del 2018 y el 27 de febrero del 2021. El archivo consultado tiene 406 451 filas. El archivo cvs se guarda en la misma carpeta donde se guardó el proyecto de R. Al abrirlo con Excel, se visualiza la siguiente información:

Identificador	Fecha reporte	Localidad	Barrio	Upz	Dirección	Tipo de afectación
4862293	1/01/2018	11 Suba	TEJARES DEL NORTE	17 - SAN JOSE DE BAVARIA	KR 49 185 38	Caída de árbol
4862292	1/01/2018	14 Los MÃ¡rtires	SANTA ISABEL SUR	37 - SANTA ISABEL	KR 29B BIS 0 64	Incendio Estructuras - Conato Estructural

4862291	1/01/2018	18 Rafael Uribe Uribe	DIANA TURBAY	55 - DIANA TURBAY	CL 48U SUR 1F 17	Accidente de tránsito
4862290	1/01/2018	11 Suba	URBANIZACION CORDOBA NIZA	24 - NIZA	KR 57 125B 73	Caída de árbol
4862289	1/01/2018	11 Suba	N/A	N/A	N/A	Conato de incendio
4862288	1/01/2018	9 FontibÃ³n	CIUDAD SALITRE OCCIDENTAL	110 - CIUDAD SALITRE OCCIDENTAL	KR 68C 22B 71	Caída de árbol
4862287	1/01/2018	15 Antonio NariÃ±o	CIUDAD BERNA	35 - CIUDAD JARDIN	CL 8 SUR 11A	Caída de árbol
4862286	1/01/2018	1 UsaquÃ©n	SAN PATRICIO	16 - SANTA BARBARA	CL 116 22 05	Incendio Estructuras - Conato Es-tructural
4862285	1/01/2018	2 Chapinero	SAN LUIS - ALTOS DEL CABO	89 - SAN ISIDRO - PATIOS	TV 5C ESTE 96A 89	Fenómeno de Remoción en Masa
4862284	1/01/2018	2 Chapinero	ANTIGUO COUNTRY	97 - CHICO LAGO	CL 87 19B	Caída de árbol
4862283	1/01/2018	N/A	N/A	N/A	N/A	Accidente de tránsito
4862282	1/01/2018	10 Engativa	LA ESTRADA	26 - LAS FERIAS	KR 70 66 14	Enfermedad o Traumatismo
4862281	1/01/2018	16 Puente Aranda	ALCALA	41 - MUZU	CL 30 SUR 51F 14	Quemas - Basuras

Primeros 13 reportes:

Últimos 11 reportes

5373414	27/02/2021	7 Bosa	JIMENEZ DE QUESADA	85 - BOSA CENTRAL	CL 68 BIS SUR 80F 23	Incendio Estructuras
5373415	27/02/2021	12 Barrios Unidos	RIONEGRO	21 - LOS ANDES	KR 58 94B 55	Daño en redes de servicio públicos alcantarillado
5373416	27/02/2021	9 Fontibón	COFRADIA	115 - CAPELLANIA	KR 97 24	Daño en redes de servicio públicos alcantarillado
5373417	27/02/2021	9 FontibÃ³n	ATAHUALPA	75 - FONTIBON	CL 22H 113A 46	Daño en redes de servicio públicos alcantarillado
5373418	27/02/2021	18 Rafael Uribe Uribe	GRANJAS SAN PABLO	53 - MARCO FIDEL SUAREZ	KR 13B 40I SUR 44	Daño en redes de servicio públicos alcantarillado
5373419	27/02/2021	13 Teusa-quillo	QUINTA PAREDES	107 - QUINTA PAREDES	KR 47 22A 95	Rescate en Espa-cios Confinados - En Ascensores
5373420	27/02/2021	12 Barrios Unidos	METROPOLIS	22 - DOCE DE OCTUBRE	CL 79A 66 66	Caída de árbol - Pérdida de v erticalidad de árbol
5373421	27/02/2021	14 Los MÃ¡rtires	SAMPER MEN-DOZA	102 - LA SABANA	CL 23A 25 32	Daño en redes de servicio públicos energía
5373422	27/02/2021	18 Rafael Uribe Uribe	QUIROGA	39 - QUIROGA	AK 14 31B SUR 35	Daño en redes de servicio públicos acueducto
5373423	27/02/2021	11 Suba	IBERIA	24 - NIZA	KR 73A 135 20	Abejas
5373424	27/02/2021	13 Teusa-quillo	ORTEZAL - QUINTA PAREDES	N/A	AC 24 46 39	Caída de árbol - Caída de ramas

5373425	27/02/2021	11 Suba	PUENTE LAR-GO	20 - LA ALHAMBRA	CL 115 BIS 58 4	Daño en redes de servicio públicos alcantarillado
5373426	27/02/2021	11 Suba	ESTORIL	N/A	AC 100 47 28	Daño en redes de servicio públicos alcantarillado
5373427	27/02/2021	10 Engativa	LA GRANJA	30 - BOYACA REAL	CL 76 78 24	Daño en redes de servicio públicos gas
5373428	27/02/2021	12 Barrios Unidos	SIETE DE AGOSTO	98 - LOS ALCAZARES	AK 24 65 57	Daño en redes de servicio públicos acueducto
5373429	27/02/2021	19 Ciudad BolÃvar	LA ACACIA	66 - SAN FRANCISCO	KR 19D 61B SUR 59	Daño en redes de servicio públicos energía
5373430	27/02/2021	8 Kennedy	CAMPO HER-MOSO	82 - PATIO BONITO	KR 88A 36 SUR 27	Daño en redes de servicio públicos alcantarillado

Con el archivo guardado, en RStudio se crea un nuevo proyecto, se elige un nombre para el proyecto (por ejemplo, "Emergencias en Bogotá") y finalizar la operación clicando en "create project".

A continuación, se cargan las librerías con las que se van a trabajar los datos:

```
14  Cargar las librerias
15
16  ```{r}
17  library(tidyverse)
18  library(sf)
19  library(ggmap)
20  library(lubridate)
21  ```
```

El texto que inicia con ‘‘‘{r} y finaliza con ‘‘‘ se le conoce con el nombre de chunk. Se ejecuta con la flecha verde en la parte superior derecha. Al ejecutar este chunk aparece información con la descripción de las librerías que se cargaron:

```
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.3     v purrr   0.3.4

## v tibble  3.1.0     v dplyr   1.0.7

## v tidyr   1.1.3     v stringr 1.4.0

## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --

## x dplyr::filter() masks stats::filter()

## x dplyr::lag()    masks stats::lag()

library(sf)

## Linking to GEOS 3.9.0, GDAL 3.2.1, PROJ 7.2.1

library(ggmap)

## Google's Terms of Service: https://cloud.google.com/maps-platform/terms/.

## Please cite ggmap if you use it! See citation("ggmap") for details.

library(lubridate)

##

## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':

##

##     date, intersect, setdiff, union
```

Para cargar un archivo con los datos de emergencias en Bogotá se elige un nombre corto, por ejemplo, emergencias, y se usa el comando read.cvs:

```
23 ```{r}
24 emergencias <-read.csv("reporte_consulta_bitacoraemergenciasfebrero2021.csv",sep
   =";",stringsAsFactors = TRUE)
```

Para mirar el contenido de este archivo se usa el comando "summary":

```
summary(emergencias)
##  Identificador             Fecha.reporte              Localidad
##  Min.   :4862264           19/12/2018:  3608          N/A       : 85099
##  1st Qu.:5008419           14/08/2018:  2944          11 Suba   : 36725
##  Median :5134204           17/07/2018:  2855          10 Engativa: 33999
##  Mean   :5131610           09/11/2018:  2713          8 Kennedy : 33404
##  3rd Qu.:5257506           24/07/2018:  2692          1 UsaquÃ©n : 27151
##  Max.   :5373430           23/07/2018:  2665          7 Bosa    : 21202
##                 (Other)   :388973   (Other)    :168870
##                    Barrio                                  Upz
##  N/A                : 44541   N/A                          : 94143
##  SAN JOSE           :  2050   28 - EL RINCON               :  8316
##  A.S.D.             :  1683   98 - LOS ALCAZARES           :  7393
##  RICAURTE           :  1668   102 - LA SABANA              :  7277
##  LAS NIEVES         :  1626   97 - CHICO LAGO              :  6930
##  SANTA BARBARA OCCIDENTAL:  1477   85 - BOSA CENTRAL :  6765
##  (Other)            :353405   (Other)                      :275626
##        Dirección                                                  Tipo.de.afectación
##  N/A                        : 4768   Accidente de tránsito                  :202306
##  AV BOYACA   :               1766   Enfermedad o Traumatismo               :128198
##  AUTOPISTA SUR:              1442   Animales Peligrosos                    :  7338
##  AV BOYACA 80 :               911   Abejas                                 :  5637
##  CL 80       :                909   Daño en redes de servicio públicos energía:       5485
##  CL 26       :                793   Daño en redes de servicio públicos gas     :      5399
##  (Other)                    :395861   (Other)                                : 52087
```

Con el comando "head" se puede revisar el nombre que tiene cada columna en el archivo:

```{r}
head(emergencias)
```

```
##   Identificador Fecha.reporte        Localidad                  Barrio
## 1       4862293    01/01/2018          11 Suba                TEJARES DEL NORTE
## 2       4862292    01/01/2018  14 Los MÃ¡rtires            SANTA ISABEL SUR
## 3       4862291    01/01/2018  18 Rafael Uribe Uribe        DIANA TURBAY
## 4       4862290    01/01/2018          11 Suba                URBANIZACION CORDOBA NIZA
## 5       4862289    01/01/2018          11 Suba                N/A
## 6       4862288    01/01/2018           9 FontibÃ³n CIUDAD SALITRE OCCIDENTAL
##                  Upz                                    Dirección
## 1  17 - SAN JOSE DE BAVARIA                          KR 49 185 38
## 2  37 - SANTA ISABEL                                 KR 29B BIS 0 64
## 3  55 - DIANA TURBAY                                 CL 48U SUR 1F 17
## 4  24 - NIZA                                         KR 57 125B 73
## 5  N/A                N/A
## 6  110 - CIUDAD SALITRE OCCIDENTAL                   KR 68C 22B 71
##                  Tipo.de.afectación
## 1                      Caida de árbol
## 2 Incendio Estructuras - Conato Estructural
## 3                   Accidente de tránsito
## 4                      Caida de árbol
## 5                    Conato de incendio
## 6                      Caida de árbol
```

Análisis temporal

Revisamos qué formato tiene la variable de fecha usando el comando "str (emergencias)":

```{r}
str(emergencias)
```

```
## 'data.frame':    406450 obs. of  7 variables:
## $ Identificador     : int  4862293 4862292 4862291 4862290 4862289 4862288 4862287 4862286 4862285 4862284 ...
## $ Fecha.reporte     : Factor w/ 1135 levels "01/01/2018","01/01/2019",..: 1 1 1 1 1 1 1 1 1 1 ...
## $ Localidad         : Factor w/ 28 levels "1 UsaquÃ©n","1 Usaquen",..: 4 7 13 4 4 25 9 1 16 16 ...
## $ Barrio            : Factor w/ 4706 levels "#¿NOMBRE?","",MADELENA",..: 4019 3793 1149 4222 2780 953 908 3697 3671 200 ...
## $ Upz               : Factor w/ 124 levels "1 - PASEO DE LOS LIBERTADORES",..: 31 53 72 39 124 17 51 30 111 121 ...
## $ Dirección         : Factor w/ 182009 levels "\037AV BOYACA 17",..: 136375 129093 56657 140887 174985 145674 84929 17711 178741 87829 ...
## $ Tipo.de.afectación: Factor w/ 127 levels "Abejas","Abejas - Ataque de Abejas a Personas",..: 18 74 6 18 23 18 18 74 68 18 ...
```

Vamos a cambiar la columna de fecha que es un factor a un formato Date, de fecha, se asigna un nuevo nombre al archivo, por ejemplo, "emergencias 1" y a continuación se visualiza el formato de la información del nuevo archivo:

```{r}
emergencias1 <- emergencias %>%
  mutate(Fecha.reporte=dmy(Fecha.reporte))
```

```{r}
str(emergencias1)
```

```
## 'data.frame':    406450 obs. of  7 variables:
## $ Identificador     : int  4862293 4862292 4862291 4862290 4862289 4862288 4862287 4862286 4862285 4862284 ...
## $ Fecha.reporte     : Date, format: "2018-01-01" "2018-01-01" ...
## $ Localidad         : Factor w/ 28 levels "1 UsaquÃ©n","1 Usaquen",..: 4 7 13 4 4 25 9 1 16 16 ...
## $ Barrio            : Factor w/ 4706 levels "#¿NOMBRE?","",MADELENA",..: 4019 3793 1149 4222 2780 953 908 3697 3671 200 ...
## $ Upz               : Factor w/ 124 levels "1 - PASEO DE LOS LIBERTADORES",..: 31 53 72 39 124 17 51 30 111 121 ...
## $ Dirección         : Factor w/ 182009 levels "\037AV BOYACA 17",..: 136375 129093 56657 140887 174985 145674 84929 17711 178741 87829 ...
## $ Tipo.de.afectación: Factor w/ 127 levels "Abejas","Abejas - Ataque de Abejas a Personas",..: 18 74 6 18 23 18 18 74 68 18 ...
```

Efectivamente, ahora el campo fecha dejó de ser factor y pasó a ser "Date". Y ya estamos en condiciones de operar con fechas. Por ejemplo, podríamos ver hace cuántos días fue cada emergencia teniendo en cuenta la fecha de hoy:

```
```{r}
emergencias1 <- emergencias1 %>%
  mutate(tiempo=today()-Fecha.reporte)
```
```

Para algunos análisis temporales esta función puede ser de utilidad.

Verificamos los encabezados de las columnas con el comando "head(emergencias1)":

```
```{r}
head(emergencias1)
```
```

```
##   Identificador Fecha.reporte      Localidad                  Barrio
## 1       4862293     2018-01-01      11 Suba                    TEJARES DEL NORTE
## 2       4862292     2018-01-01      14 Los MÃ¡rtires           SANTA ISABEL SUR
## 3       4862291     2018-01-01      18 Rafael Uribe Uribe      DIANA TURBAY
## 4       4862290     2018-01-01      11 Suba                    URBANIZACION CORDOBA NIZA
## 5       4862289     2018-01-01      11 Suba                    N/A
## 6       4862288     2018-01-01      9 FontibÃ³n                CIUDAD SALITRE OCCIDENTAL
##            Upz                      Dirección
## 1       17 - SAN JOSE DE BAVARIA    KR 49 185 38
## 2       37 - SANTA ISABEL           KR 29B BIS 0 64
## 3       55 - DIANA TURBAY           CL 48U SUR 1F 17
## 4       24 - NIZA                   KR 57 125B 73
## 5       N/A                         N/A
## 6 110 - CIUDAD SALITRE OCCIDENTAL   KR 68C 22B 71
##            Tipo.de.afectación                    tiempo
## 1       Caida de árbol                           1675 days
## 2 Incendio Estructuras - Conato Estructural      1675 days
## 3       Accidente de tránsito                    1675 days
## 4       Caida de árbol                           1675 days
## 5       Conato de incendio                       1675 days
## 6       Caida de árbol                           1675 days
```

Y una vista general del contenido del archivo con el comando "summary (emergencias1)". Esta función se utiliza cuando estamos conociendo el contenido de un set de datos que nos dará un resumen en forma de estadísticas descriptivas para las variables numéricas (cuartiles y mediana) y un vistazo a las categorías más representadas para los factores.

```{r}
summary(emergencias1)
```

```
##  Identificador          Fecha.reporte                Localidad
##  Min.   :4862264        Min.   :2018-01-01           N/A       : 85099
##  1st Qu.:5008419        1st Qu.:2018-06-27           11 Suba    : 36725
##  Median :5134204        Median :2018-09-12           10 Engativa: 33999
##  Mean   :5131610        Mean   :2018-10-26           8 Kennedy  : 33404
##  3rd Qu.:5257506        3rd Qu.:2019-02-05           1 UsaquÃ©n : 27151
##  Max.   :5373430        Max.   :2021-02-27           7 Bosa     : 21202
##                                                      (Other)    :168870
##                  Barrio                          Upz
##  N/A                : 44541   N/A                    : 94143
##  SAN JOSE           :  2050   28 - EL RINCON         :  8316
##  A.S.D.             :  1683   98 - LOS ALCAZARES     :  7393
##  RICAURTE           :  1668   102 - LA SABANA        :  7277
##  LAS NIEVES         :  1626   97 - CHICO LAGO        :  6930
##  SANTA BARBARA OCCIDENTAL:  1477   85 - BOSA CENTRAL :  6765
```

```
##  (Other)            :353405   (Other)                 :275626
##         Dirección                                              Tipo.de.afectación
##  N/A         :  4768   Accidente de tránsito                  :202306
##  AV BOYACA   :  1766   Enfermedad o Traumatismo               :128198
##  AUTOPISTA SUR:  1442   Animales Peligrosos                    :  7338
##  AV BOYACA 80 :   911   Abejas                                 :  5637
##  CL 80 :   909   Daño en redes de servicio públicos energía:  5485
##  CL 26       :   793   Daño en redes de servicio públicos gas  :  5399
##  (Other)     :395861   (Other)                                :  52087
##    tiempo
##  Length:406450
##  Class :difftime
##  Mode  :numeric
##
##
##
##
```

En estadística es muy común el uso de histogramas o diagramas de barras para la representación de datos. El histograma es un diagrama de barras adyacentes en una línea de base donde el eje vertical representa la frecuencia y el eje horizontal, el rango. En el histograma no existe separación entre las barras. Los gráficos de barras se emplean habitualmente para ver la evolución en el tiempo de una magnitud concreta o comparar magnitudes de varias categorías (Bembibre, 2009).

Hagamos un histograma para entender la distribución de la variable numérica que creamos:

```r
```{r}
ggplot(emergencias1)+
  geom_histogram(aes(x=Fecha.reporte))

```
```

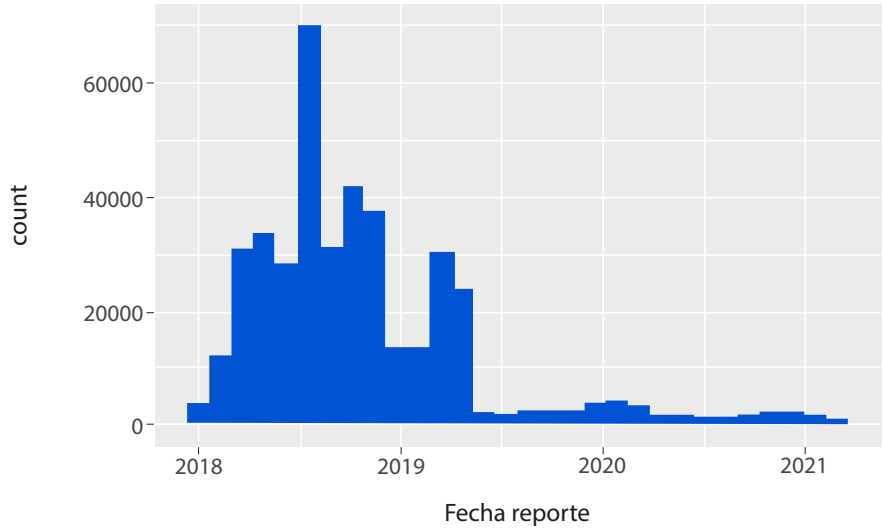


Figura 3. Ejemplo de un histograma con la distribución de las emergencias reportadas en la ciudad de Bogotá entre los años 2018 y 2021

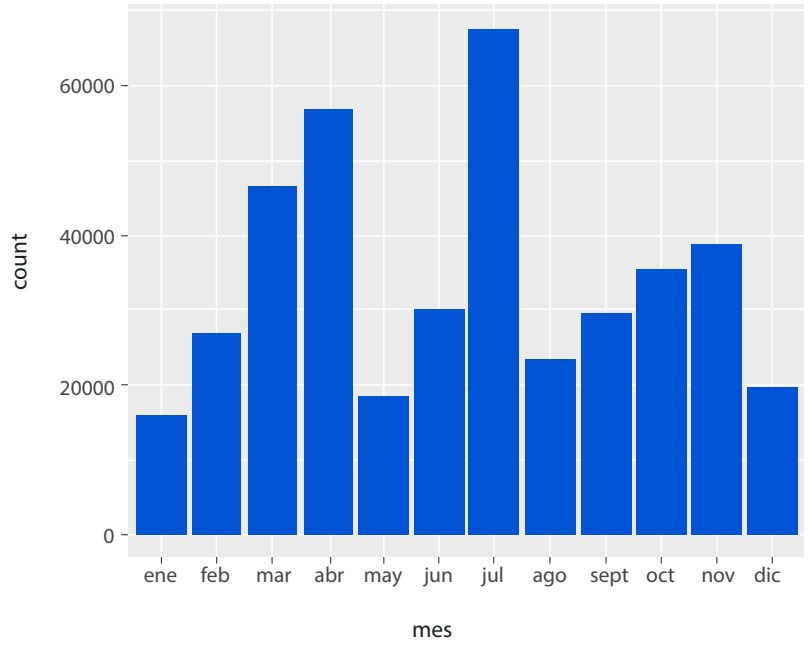
En está gráfica se puede notar la disminución en el reporte de emergencias durante el confinamiento producido por la pandemia del covid 19.

También se puede hacer un gráfico que muestre cómo es el comportamiento de los reportes de emergencia según el mes del año usando los siguientes comandos:

```
emergencias1 <- emergencias1 %>%
mutate(mes=month(Fecha.reporte, label = TRUE))
ggplot(emergencias1) +
geom_bar(aes(x = mes))
```

```r
emergencias1 <- emergencias1 %>%
  mutate(mes=month(Fecha.reporte, label = TRUE))
```

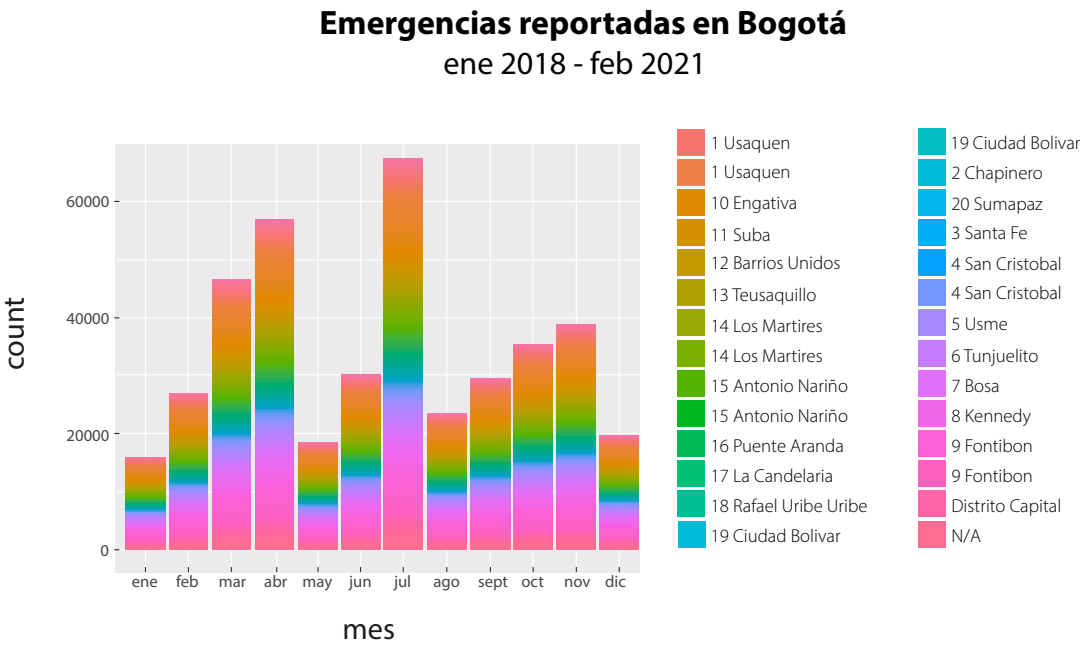
```r
ggplot(emergencias1) +
    geom_bar(aes(x = mes))
```



► **Figura 4.** Gráfico de barras que indica la cantidad de emergencias reportadas según el mes del año en el periodo comprendido entre enero de 2018 y febrero de 2021.

Con el comando "labs" se pueden incluir atributos de la gráfica como título, subtítulo, nombre de los ejes, fuente de los datos, entre otras.

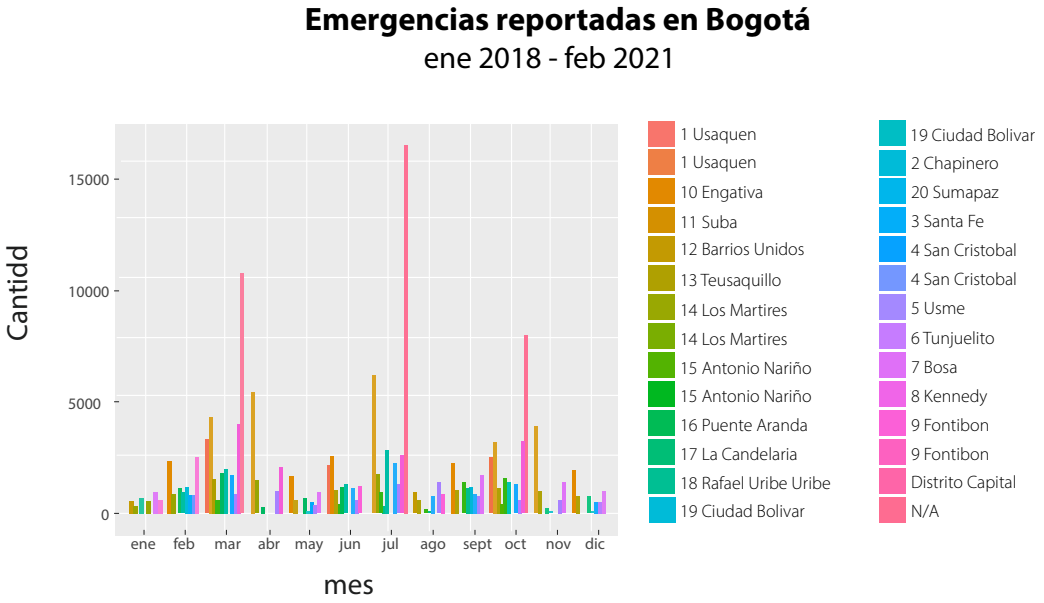
```r
ggplot(emergencias1) +
    geom_bar(aes(x = mes, fill=Localidad)) +
    labs(title = "Emergencias reportadas en Bogotá ",
         subtitle = "ene 2018 - feb 2021",
         x = "Mes",
         y = "Cantidad",
         fill = "Tipo de Emergencia",
         caption = "Fuente: Alcadia Bogotá")
```



► **Figura 5**. Gráfico de barras donde se representa la cantidad de reportes de emergencia durante el periodo comprendido entre enero de 2018 y febrero de 2021 discriminado por localidades en la ciudad de Bogotá

Y también probemos el gráfico sin apilar usando el comando "position = "dodge"":

```r
ggplot(emergencias1) +
    geom_bar(aes(x = mes, fill=Localidad), position = "dodge") +
    labs(title = "Emergencias reportadas en Bogotá",
         subtitle = "ene 2018 - feb 2021",
         x = "Mes",
         y = "Cantidad",
         fill = "Tipo de Emergencia",
         caption = "Fuente: Alcadia Bogotá")
```



► **Figura 6**. Gráfico de barras sin apilar donde se representa la cantidad de reportes de emergencia durante enero de 2018 y febrero de 2021 discriminado por localidades en la ciudad de Bogotá

Las ventajas de un buen gráfico están en que puede captar la atención del lector, presentar los datos de una manera sencilla y fácil de leer destacando tendencias y diferencias.

Los datos se pueden clasificar en dos grupos: cuantitativos y cualitativos. Cuantitativos se refieren a cantidades o valores numéricos. Dentro de los datos cuantitativos estos

pueden ser discretos, si toman valores enteros, o continuos, si pueden tomar cualquier valor dentro de un intervalo. Cualitativos se refieren a cualidades o modalidades que no pueden expresarse numéricamente. Los datos cualitativos son ordinales, si siguen un orden o secuencia, o categóricos, si no siguen ningún orden (Instituto Nacional de Estadística, s. f.).

Ejercicio 2. Proyecto en R: Medellín

Sitios de interés cultural

El primer paso es plantearse una pregunta por resolver a través del análisis de unos datos. En este ejercicio, la pregunta está relacionada con los sitios de interés de la ciudad de Medellín y la disponibilidad de rutas de transporte público para llegar a ellos.

El siguiente paso es buscar información en las bases de datos abiertos. Para este ejercicio se consultó el siguiente link:

https://geomedellin-m-medellin.opendata.arcgis.com/datasets/M-Medellin::patrimonio-bienes-de-interes-cultural-bic/explore?location=6.288508%2C-75.476202%2C10.75

Se descarga el archivo cvs y se abre en Excel. Es importante guardar en la misma carpeta donde se tiene el proyecto para luego poder trabajar estos datos en R. Los datos deben estar en la misma carpeta donde se encuentra alojado el archivo para que se puedan ejecutar en el "script".

Ciudad de Medellín

Primero cargamos las librerías que vamos a usar:

```r
```{r}
library(tidyverse)
library(sf)
```
```

```
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 -

## v ggplot2 3.3.3     v  abí   0.3.4

## v tibble  3.1.0     v dplyr   1.0.7

## v tidyr   1.1.3     v stringr 1.4.0

## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --

## x dplyr::filter() masks stats::filter()

## x dplyr::lag()    masks stats::lag()

library(sf)

## Linking to GEOS 3.9.0, GDAL 3.2.1, PROJ 7.2.1
```

Los datos se pueden encontrar en diferentes formatos como CSV que es un archivo separado por comas o como el que se trabaja a continuación JSON. Un archivo JSON es un documento digital creado en este lenguaje que almacena información organizada, con el fin de hacer más simple su búsqueda y acceso. La ventaja de este formato es que permite obtener código legible para las personas con nombres y valores que funcionan como indicadores de la información que contienen (Coppola, 2022).

Cargamos un *dataset* espacial que contiene datos de la ciudad de Medellín, situada en Colombia, con información de bienes de interés cultural.

```r
Medellin_BIC <- st_read("Patrimonio_-_Zona_de_Influencia_Bienes_de_Interes_Cultural.geojson")
```

```r
summary(Medellin_BIC)
```

```
## Reading layer `Patrimonio_-_Zona_de_Influencia_Bienes_de_Interes_Cultural' from data source `C:\Users\MARGARITA\OneDrive\Documentos\Margarita\Politicas pÃºblicas\Datos de ciudades\Patrimonio_-_Zona_de_Influencia_Bienes_de_Interes_Cultural.geojson' using driver `GeoJSON'

## Simple feature collection with 18 features and 8 fields

## Geometry type: POLYGON

## Dimension:     XY

## Bounding box:  xmin: -75.59701 ymin: 6.208619 xmax: -75.55668 ymax: 6.279218

## Geodetic CRS:  WGS 84

summary(Medellin_BIC)

##    OBJECTID         GRUPO            SUBGRUPO           TIPO

## Min.   : 1.00    Length:18          Length:18          Length:18

## 1st Qu.: 5.25    Class :character   Class :character   Class :character

## Median : 9.50    Mode  :character   Mode  :character   Mode  :character

## Mean   : 9.50

## 3rd Qu.:13.75

## Max.   :18.00

##    NOMBRE          DIRECCION          SHAPEAREA            SHAPELEN

## Length:18        Length:18          Min.   :  5392       Min.   : 286.6
```

```
## Class :character        Class :character        1st Qu.:  23645        1st Qu.: 747.2
## Mode  :character        Mode  :character        Median :  67242        Median :1215.7
##                                                 Mean   : 252934        Mean   :1827.8
##                                                 3rd Qu.: 129910        3rd Qu.:1674.5
##                                                 Max.   :1760133        Max.   :6868.6
##          geometry
## POLYGON      :18
## epsg:4326    : 0
## +proj=long...: 0
##
##
##
```

En R, el argumento "stringsAsFactors" se utiliza para indicar si las cadenas de caracteres en un data frame deben ser tratadas como factores o como cadenas de caracteres simples. Por defecto, el valor de este argumento es TRUE, lo que significa que las cadenas de caracteres se convierten en factores. Si se establece FALSE, las cadenas de caracteres se mantiene como tales ("Introduction to importing data in R", s. f.).

Vamos a cargar las variables como factor:

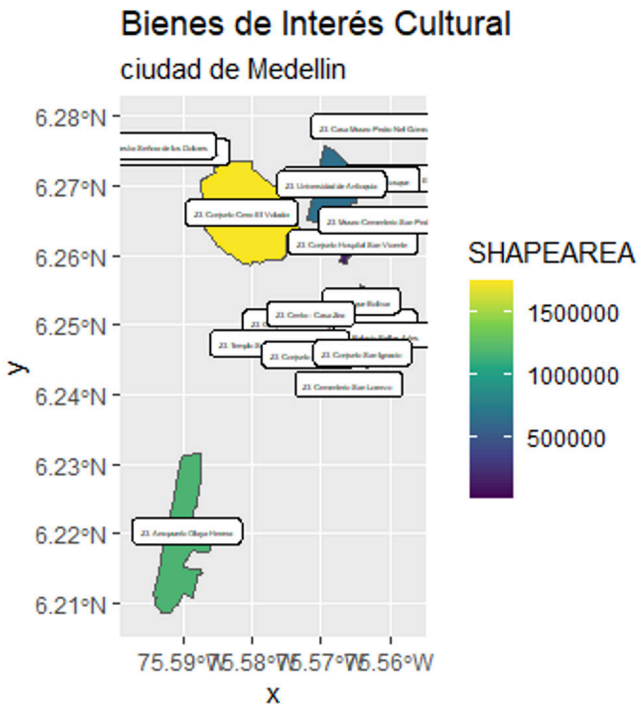
```r
Medellin_BIC <- st_read("Patrimonio_-_Zona_de_Influencia_Bienes_de_Interes_Cultural.geojson",stringsAsFactors=TRUE)
```

```
## Reading layer `Patrimonio_-_Zona_de_Influencia_Bienes_de_Interes_Cultural' from data source `C:\Users\MARGARITA\OneDrive\Documentos\Margarita\Politicas pÃºblicas\Datos de ciudades\Patrimonio_-_Zona_de_Influencia_Bienes_de_Interes_Cultural.geojson' using driver `GeoJSON'

## Simple feature collection with 18 features and 8 fields

## Geometry type: POLYGON

## Dimension:     XY

## Bounding box:  xmin: -75.59701 ymin: 6.208619 xmax: -75.55668 ymax: 6.279218

## Geodetic CRS:  WGS 84

summary(Medellin_BIC)

##     OBJECTID                     GRUPO                                SUBGRUPO

##  Min.   : 1.00                   Zona de influencia Arqueólogica: 1   N.A:18

##  1st Qu.: 5.25                   Zona de influencia Urbanística :17

##  Median : 9.50

##  Mean   : 9.50

##  3rd Qu.:13.75

##  Max.   :18.00

##

##                                TIPO

##  N.A                              : 1

##  Sector con tratamiento de Conservación Nivel 3:17

##
```

```
##
##
##
##
##                      NOMBRE                  DIRECCION                SHAPEAREA
## Z.I. Aeropuerto Olaya Herrera  : 1       NA's:18   Min.   :          5392
## Z.I. Casa Museo Pedro Nel Gómez: 1    1st Qu.:                      23645
## Z.I. Cementerio San Lorenzo    : 1       Median :                    67242
## Z.I. Centro - Casa Zea         : 1       Mean   :                    252934
## Z.I. Conjunto Cerro El Volador : 1       3rd Qu.:                    129910
## Z.I. Conjunto Guayaquil        : 1       Max.   :                    1760133
## (Other)                        :12
##    SHAPELEN             geometry
## Min.   : 286.6   POLYGON     :18
## 1st Qu.: 747.2   epsg:4326   : 0
## Median :1215.7   +proj=long...: 0
## Mean   :1827.8
## 3rd Qu.:1674.5
## Max.   :6868.6
##
```

Con la función ggplot que permite realizar un gráfico como se muestra a continuación:

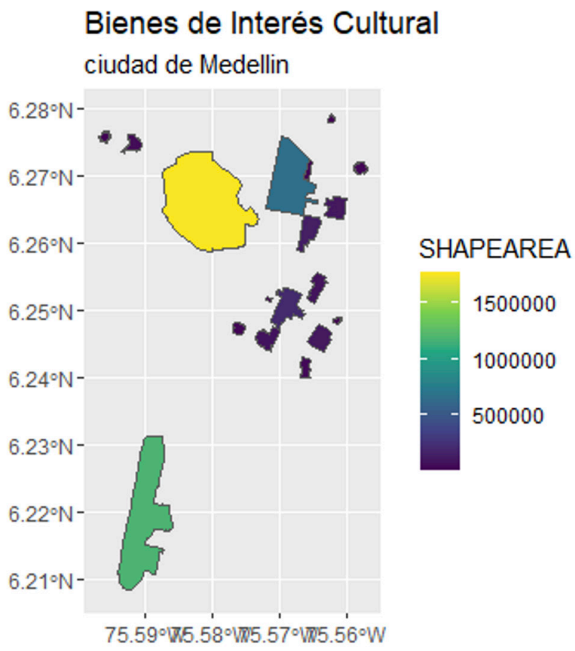
```
ggplot(Medellin_BIC)+
 geom_sf()+
 geom_sf(aes(fill=SHAPEAREA))+
 labs(title="Bienes de Interés Cultural ",
     subtitle="ciudad de Medellin")+
 scale_fill_viridis_c()+
 geom_sf_label(aes(label=NOMBRE), size=1.1)
## Warning in st_point_on_surface.sfc(sf::st_zm(x)): st_point_on_surface may not
## give correct results for longitude/latitude data
```



► **Figura 7.** Visualización de los sitios de interés cultural, ciudad de Medellín

Como las etiquetas en la figura 3 ocupan mucho espacio en el mapa las vamos a omitir.

```
ggplot(Medellin_BIC) + geom_sf() + geom_sf(aes(fill = SHAPEAREA)) + labs(title = "Bienes de Interés Cultural ", subtitle = "ciudad de Medellin") +  scale_fill_viridis_c()
```



► **Figura 8**. Visualización de los sitios de interés cultural, ciudad de Medellín, sin etiquetas

Ahora le vamos a agregar al mapa las vías de transporte público usando otro data set.

```
vias<- st_read("Corredores_para_Transporte_de_Pasajeros.geojson",stringsAsFactors=TRUE)

## Reading layer `Corredores_para_Transporte_de_Pasajeros' from data source `C:\Users\MARGARITA\OneDrive\Documentos\Margarita\Politicas pÃºblicas\Datos de ciudades\Corredores_para_Transporte_de_Pasajeros.geojson' using driver `GeoJSON'

## Simple feature collection with 94 features and 7 fields

## Geometry type: MULTILINESTRING
```

```
## Dimension:     XY

## Bounding box:  xmin: -75.72896 ymin: 6.074753 xmax: -75.4407 ymax: 6.368562

## Geodetic CRS:  WGS 84

summary( abí)

##     OBJECTID                    TIPO_SISTEMA                    LINEA

##  Min.   : 1.00                  Min.   :1.000                   Sin definir:82

##  1st Qu.:24.25                  1st Qu.:5.000                   Línea 1    : 2

##  Median :47.50                  Median :5.000                   Línea 2    : 2

##  Mean   :47.50                  Mean   :4.468                   Línea A    : 1

##  3rd Qu.:70.75                  3rd Qu.:5.000                   Línea B    : 1

##  Max.   :94.00                  Max.   :5.000                   Línea C    : 1

##                                 (Other)    : 5

##                          NOMBRE_TRAMO                          NOMBRE

##  Corredor Girardot - Mont y Velarde    : 4                     Carrera 50: 4

##  Corredor Av. 34                       : 3                     Carrera 80 - 81: 3

##  Corredor Av. Carabobo                 : 3                     Girardot - Mont y Velarde : 3

##  Corredor Longitudinal Oriental        : 3                     Longitudinal Oriental     : 3

##  Cable Palmitas                        : 2                     Palmitas                  : 3

##  Corredor ''U'' Corta                  : 2                     Av. Carabobo              : 2

##  (Other)                               :77                     (Other)                   :76

##     ESTADO        SHAPELEN                  geometry

##  Min.   :1.000   Min.   :  77.48   MULTILINESTRING:94
```

```
##  1st Qu.:3.000   1st Qu.:  960.78   epsg:4326     : 0
##  Median :3.000   Median : 1815.60   +proj=long...  : 0
##  Mean   :2.894   Mean   : 3811.79
##  3rd Qu.:3.000   3rd Qu.: 4918.88
##  Max.   :5.000   Max.   :36411.39
##
```

Usamos nuevamente el comando ggplot para generar un gráfico que incluya las vías de transporte público.

```
ggplot()+
 geom_sf(data=Medellin_BIC, aes(fill=SHAPEAREA))+
 geom_sf(data=vias)
```

► **Figura 9.** Visualización del mapa con las vías de transporte público, ciudad de Medellín

Al colocar el mapa de las vías se puede notar que existen conexiones de transporte público para acceder a los sitios catalogados como bienes de interés cultural.

De acuerdo con la situación problema que se haya planteado en un inicio se puede buscar información en los diferentes portales abiertos que existen para ser analizados y se pueden combinar estos datos de tal forma que permita tener argumentos para facilitar la toma de decisiones. Las entidades encargadas de la Administración Pública utilizan esta herramienta dentro de su gestión.

Ejercicio 3. Proyecto en R: Medellín

Patrimonio

Primero cargamos la librería tidyverse:

```
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.3     v purrr   0.3.4

## v tibble  3.1.0     v dplyr   1.0.7

## v tidyr   1.1.3     v stringr 1.4.0

## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --

## x dplyr::filter() masks stats::filter()

## x dplyr::lag()    masks stats::lag()
```

Luego se carga la librería sf:

Esta librería permite trabajar con el modelo de geometrías de *características simples* (o rasgos simples) es un estándar (ISO 19125) desarrollado por el Open Geospatial Consortium (OGC) para formas geográficas vectoriales, que ha sido adoptado por gran cantidad de *software* geográfico (entre otros por GeoJSON, ArcGIS, QGIS, PostGIS, MySQL Spatial Extensions, Microsoft SQL

Server...). Como ya se comentó, este tipo de datos espaciales se está implementado en R en el paquete sf (Fernández y Cotos, 2022).

```
library(sf)

## Linking to GEOS 3.9.0, GDAL 3.2.1, PROJ 7.2.1

Medellin_BIC <- st_read("Patrimonio_-_Zona_de_Influencia_Bienes_de_Interes_Cultural.geojson")

## Reading layer `Patrimonio_-_Zona_de_Influencia_Bienes_de_Interes_Cultural' from data source `C:\Users\MARGARITA\OneDrive\Documentos\Margarita\Politicas pÃºblicas\Datos de ciudades\Patrimonio_-_Zona_de_Influencia_Bienes_de_Interes_Cultural.geojson' using driver `GeoJSON'

## Simple feature collection with 18 features and 8 fields

## Geometry type: POLYGON

## Dimension:     XY

## Bounding box:  xmin: -75.59701 ymin: 6.208619 xmax: -75.55668 ymax: 6.279218

## Geodetic CRS:  WGS 84

summary(Medellin_BIC)

##     OBJECTID                 GRUPO                SUBGRUPO             TIPO

##  Min.   : 1.00               Length:18            Length:18            Length:18

##  1st Qu.: 5.25               Class :character     Class :character     Class :character

##  Median : 9.50               Mode  :character     Mode  :character     Mode  :character

##  Mean   : 9.50

##  3rd Qu.:13.75

##  Max.   :18.00
```

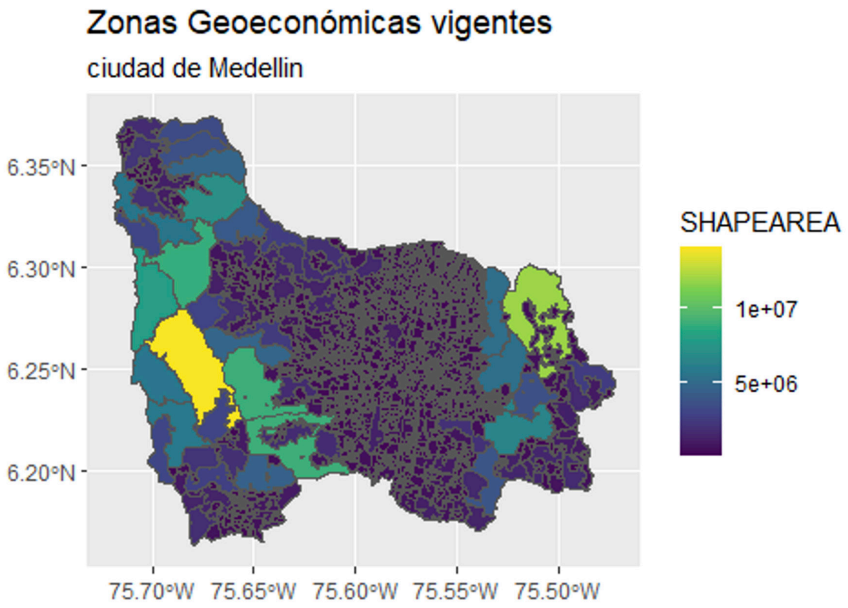
```
##    NOMBRE             DIRECCION          SHAPEAREA        SHAPELEN
## Length:18          Length:18          Min.   :  5392    Min.   : 286.6
## Class :character   Class :character   1st Qu.:  23645    1st Qu.: 747.2
## Mode  :character   Mode  :character   Median :  67242    Median :1215.7
##                                       Mean   : 252934    Mean   :1827.8
##                                       3rd Qu.: 129910    3rd Qu.:1674.5
##                                       Max.   :1760133    Max.   :6868.6
##        geometry
## POLYGON     :18
## epsg:4326   : 0
## +proj=long...: 0
##
##
##
```

```
economia <- st_read("Zonas_Geoeconómicas_Vigentes_-_Acuerdo_48_de_2014.geojson",
stringsAsFactors=TRUE)
```

```
## Reading layer `Zonas_Geoecon%C3%B3micas_Vigentes_-_Acuerdo_48_de_2014' from data
source `C:\Users\MARGARITA\OneDrive\Documentos\Margarita\Politicas pÃºblicas\Datos de
ciudades\Zonas_GeoeconÃ³micas_Vigentes_-_Acuerdo_48_de_2014.geojson' using driver
`GeoJSON'
## Simple feature collection with 3406 features and 7 fields
## Geometry type: MULTIPOLYGON
```

```
## Dimension:    XY
## Bounding box:  xmin: -75.71922 ymin: 6.163198 xmax: -75.47185 ymax: 6.373804
## Geodetic CRS:  WGS 84
summary(economia)
##    OBJECTID                 COMUNA             VALOR_M2        COD_VALOR
##  Min.   :3521   10             : 241   Min.   :      3567   Min.   :196.0
##  1st Qu.:4372   16             : 229   1st Qu.:    259763   1st Qu.:420.0
##  Median :5224   07             : 227   Median :    460317   Median :450.0
##  Mean   :5224   90             : 225   Mean   :    780858   Mean   :446.9
##  3rd Qu.:6075   08             : 208   3rd Qu.:   1085275   3rd Qu.:494.0
##  Max.   :6926   (Other):2274   Max.   :15097837   Max.   :633.0
##                 NA's   :   2
##    VIGENCIA               SHAPEAREA                SHAPELEN          geometry
##  Min.   :2021          Min.   :      7   Min.   :    11.0   MULTIPOLYGON :3406
##  1st Qu.:2021          1st Qu.:   1307   1st Qu.:   161.7   epsg:4326    :   0
##  Median :2021          Median :   5826   Median :   392.0   +proj=long...:   0
##  Mean   :2021          Mean   : 110498   Mean   :  1236.8
##  3rd Qu.:2021          3rd Qu.:  38479   3rd Qu.:  1350.9
##  Max.   :2021          Max.   :14135150   Max.   :  48713.7
##
ggplot(economia)+
```

```
geom_sf()+
geom_sf(aes(fill=SHAPEAREA))+
labs(title="Zonas Geoeconómicas vigentes ",
    subtitle="ciudad de Medellin")+
scale_fill_viridis_c()
```



► **Figura 10.** Mapa generado en R para las zonas geoeconómicas de la ciudad de Medellín

```
zonas_fisicas <- st_read("Zonas_Homogeneas_Fisicas_Urbanas.geojson",stringsAsFactors=-
TRUE)

## Reading layer `Zonas_Homogeneas_Fisicas_Urbanas' from data source `C:\Users\MARGA-
RITA\OneDrive\Documentos\Margarita\Politicas pÃºblicas\Datos de ciudades\Zonas_Homo-
geneas_Fisicas_Urbanas.geojson' using driver `GeoJSON'
```

```
## Simple feature collection with 42391 features and 26 fields (with 15 geometries empty)
## Geometry type: MULTIPOLYGON
## Dimension:     XY, XYZ
## Bounding box:  xmin: -75.66447 ymin: 6.165494 xmax: -75.52479 ymax: 6.312832
## z_range:       zmin: 0 zmax: 0
## Geodetic CRS:  WGS 84
summary(zonas_fisicas)
##     OBJECTID                    CODIGO_ZONA_FISICA                  COMUNA
##  Min.   :    1                  108472501431457:  228   14          : 8250
##  1st Qu.:10598                  108472502431457:  220   11          : 3114
##  Median :21196                  102472501431457:  140   16          : 2945
##  Mean   :21196                  108322120431422:  137   08          : 2677
##  3rd Qu.:31794                  108112502431415:  124   07          : 2479
##  Max.   :42391                  108112502451415:  110   09          : 2439
##                (Other)       :41432   (Other):20487
##     VIGENCIA                          CLASE_SUELO              CODIGO_CS
##  Min.   :2019-12-31 19:00:00          Expansión   :  333          1:42058
##  1st Qu.:2019-12-31 19:00:00          Suelo Urbano:42058        3:  333
##  Median :2019-12-31 19:00:00
##  Mean   :2020-04-07 21:46:39
##  3rd Qu.:2020-12-31 19:00:00
```

```
##  Max.   :2020-12-31 19:00:00
##
##                                              DESCRIPC_CS                  CODIGO_AP
##  Sin restricción                             :24385   41                  :17223
##  Retiros a corrientes hídricas               : 9578   31                  : 9578
##  Areas de amenaza alta                       : 3954   08                  : 7162
##  AIE - Sistema hidrografico (EEP)            : 3280   21                  : 3954
##  AIE - Red de conectividad ecologica (EEP):  645   02                    : 3280
##  Zonas con condiciones de riesgo             :  146   04                  :  645
##  (Other)                                     :  403   (Other)             :549
##                                                  DESCRIPC_AP          CODIGO_UPO
##  Residencial Predominante                        : 8546   11           :11005
##  Zonas Verdes Recreacionales                     : 4634   61           : 5563
##  Espacio Publico Proyectado                      : 4627   44           : 5432
##  Areas y corredores de media mixtura             : 3145   21           : 4696
##  Centralidades y corredores con alta intensidad: 2472   32             : 3593
##  Baja Mixtura- Residencial predominante          : 2459   51           : 2476
##  (Other)                                    :16508   (Other): 9626
##       DESCRIPC_UPO                 CODIGO_TU          CODIGO_ZGR
##  Consolidacion Nivel 5             :9005   25           :9005   0:36573
##  Consolidacion Nivel 2             :8684   22           :8684   1:  616
##  Consolidacion Nivel 3             :7383   23           :7383   2: 5202
```

```
##  Mejoramiento Integral en suelo urbano:4154   31    :4154
##  Renovación Urbana                :3541   41    :3541
##  Consolidacion Nivel 1            :3137   21    :3137
##  (Other)                          :6487   (Other):6487
##                DESCRIPC_ZGR   CODIGO_P                    DESCRIPC_P
##  Zona Receptora       : 2875   0:13326                   Ligeramente Inclinado (11-25%):10474
##  Zonas generadoras    :  616   1:13131                   Plano (0-10%)                : 8467
##  Zonas no clasificadas:36573   2: 9579                   Inclinado (26-40%)           : 5731
##  Zonas receptoras     : 2327   3: 4869                   Ligeramente Escarpado (41-60%): 4869
##                                4: 1486   Plano 0% - 7%                : 4859
##                                Empinado 14% - 60%           : 3848
##                                (Other)                      : 4143
##  CODIGO_SP                                    DESCRIPC_SP           CODIGO_MOV
##  0:   51   Completos                                  :   28   51          :23258
##  2:   28   Completos mas dos complementarios:19412   31          : 8461
##  3:11536   Completos más dos complementarios:11364   41          : 6825
##  4:30776   Completos mas un complementario  :11536   11          : 1042
##            Sin Servicios                    :   51   21     :  934
##                                             12     :  837
##                                             (Other): 1034
##    DESCRIPC_MOV              CODIGO_EV           DESCRIPC_EV     CODIGO_UA
##  Via Servicio    :19580          0:  102   Regular      :16349   13          : 6647
```

```
##  Vías Arterias   : 5674   1:   64   Bueno    :10826   12    : 5349
##  Via Colectora   : 5281   2: 9155   Malo     : 9155   57    : 3618
##  Vías de Servicio: 3680   3:16349   Bueno    : 5224   14    : 3495
##  Via Arteria     : 2787   4:16050   Excelente:  671   52    : 2945
##  Vías Colectoras : 1544   5:  671   No aplica:  102   22    : 2730
##  (Other)         : 3845             (Other)  :   64   (Other):17607
##                DESCRIPC_UA     SHAPEAREA          SHAPELEN
##  Residencial (Tipo3)   : 6452   Min.   :    0.0   Min.   :    0.0
##  Residencial (Tipo2)   : 5332   1st Qu.:   18.4   1st Qu.:   46.8
##  Espacio Publico de vias: 2963   Median :   226.9   Median :  119.3
##  Zonas Verdes          : 2842   Mean   :  2161.5   Mean   :  301.4
##  Residencial Tipo 5    : 2464   3rd Qu.:  1300.4   3rd Qu.:  271.4
##  (Other)               :22291   Max.   :922700.7   Max.   :45352.6
##  NA's                  :   47
##          geometry
##  MULTIPOLYGON  :   15
##  MULTIPOLYGON Z:42376
##  epsg:4326     :    0
##  +proj=long... :    0
##
##
##
```

Ahora vamos a cargar un archivo con los datos correspondientes a zonas de espacio público.

```
espacio_publico <- read.csv("Espacio_Publico_Existente.csv", stringsAsFactors = TRUE)

head(espacio_publico)

##   ï..OBJECTID           NOMBRE FUNCION                CBML            CATEGORIA

## 1           1 Sin InformaciÃ³n       5 01010730009  Zona Verde Recreacional

## 2           2 Sin InformaciÃ³n       5  01030730002 Zona Verde Recreacional

## 3           3 Sin InformaciÃ³n       5  01090040001 Zona Verde Recreacional

## 4           4 Sin InformaciÃ³n       5  02010010131 Zona Verde Recreacional

## 5           5 Sin InformaciÃ³n       5  03090540001 Zona Verde Recreacional

## 6           6 Sin InformaciÃ³n       5  03080240013 Zona Verde Recreacional

##   ORDEN                          DOMINIO NIVEL SUBCATEGORIA LABEL SHAPEAREA

## 1     2 Area libre del equipamiento pÃºblico    VV      <Null>       244.509

## 2     2 Area libre del equipamiento pÃºblico   BS2      <Null>      3853.746

## 3     2 Area libre del equipamiento pÃºblico    VV      <Null>       903.769

## 4     2 Area libre del equipamiento pÃºblico    VV      <Null>       271.999

## 5     2 Area libre del equipamiento pÃºblico   BS2      <Null>      4584.791

## 6     2 Area libre del equipamiento pÃºblico    VV      <Null>       689.579

##    SHAPELEN COD_BARRIO COD_COMUNA             NOM_BARRIO

## 1 142.51180       0101        01 Santo Domingo Savio No.1

## 2 501.19046       0103        01               Popular

## 3 164.36858       0109        01          Aldea Pablo VI
```

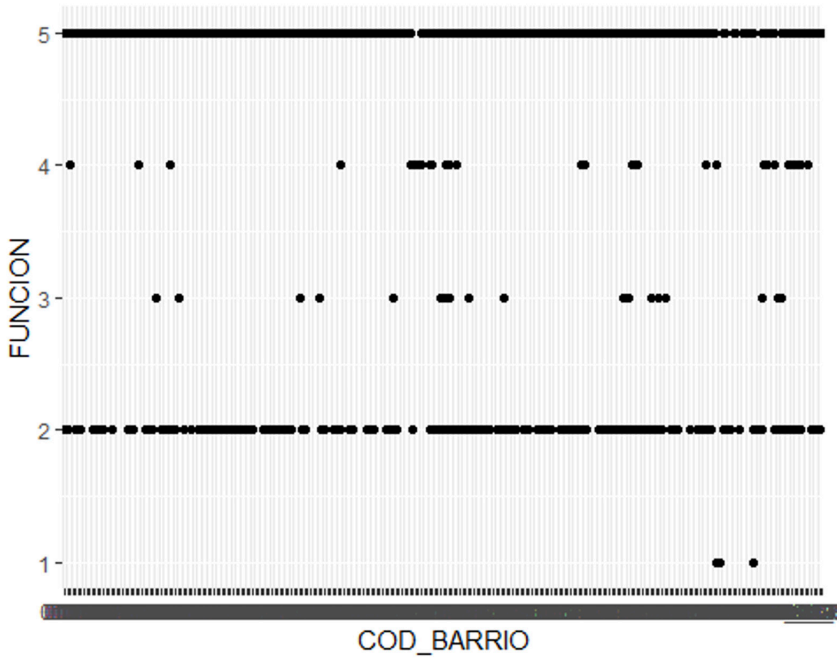
```
## 4  80.35244      0201       02              La Isla
## 5 528.63605      0309       03         Versalles No.1
## 6 233.01174      0308       03       Manrique Oriental
summary(espacio_publico)
##  ï..OBJECTID                                NOMBRE                    FUNCION
##  Min.   :  1.0   Sin InformaciÃ³n                :3131          Min.   :1.000
##  1st Qu.: 941.5   Parque                        :  52          1st Qu.:5.000
##  Median :1877.0   Zona Verde                    :  34          Median :5.000
##  Mean   :1876.8   Espacio PÃºblico Plan Parcial Pajarito:  23   Mean   :4.635
##  3rd Qu.:2812.5   Parque Infantil               :  14   3rd Qu.:5.000
##  Max.   :3842.0   Parque Lineal Quebrada La Bermejala  :  12   Max.   :5.000
##                   (Other)                       : 477
##          CBML                    CATEGORIA                    ORDEN
##  No Posee    :                  463   Ecoparque              :  54   Min.   :1.000
##  16050920001 :                   72   Mirador Panoramico     :   4   1st Qu.:2.000
##  Tiene varios:                   47   Parque CÃ¬vico          :  24   Median :2.000
##  11160010001 :                   42   Parque recreativo      : 362   Mean   :1.977
##  07030070001 :                   41   Plaza                  :  50   3rd Qu.:2.000
##  13040010001 :                   37   Zona Verde Recreacional:3249   Max.   :2.000
##  (Other)     :3041
##                        DOMINIO
```

```
##  Area libre del equipamiento pÃºblico:       117
##  Area libre privada de uso pÃºblico  :       19
##  Bienes de uso pÃºblico              :       3607
##
##
##
##
##                                              NIVEL
##  BS2                                         : 935
##  CS1                                         : 532
##  M                                           :  72
##  Nivel Barrial/ Suburbano Nivel 2:   1
##  RM                                          :  15
##  VV                                          :2125
##  ZC                                          :  63
##                                                                SUBCATEGORIA
##  <Null>                                                         :3280
##  Parque recreativo pasivo                                       : 273
##  Parque recreativo activo                                       :  89
##  Ecoparque de quebrada y otros cuerpos de agua                  :  47
##  Plazoleta                                                      :  34
##  Plazuela                                                       :   8
```

```
##  (Other)                          : 12
##                                  LABEL                                    SHAPEAREA
##                                    :3247   Min.   :                        0.0
##  EP Plan Parcial Pajarito          : 23   1st Qu.:                        133.6
##  Parque Lineal Quebrada La Bermejala: 12   Median :                        489.2
##  Parque Lineal Quebrada La Quintana : 11   Mean   :                        2732.9
##  Z.V. Urb. Mano de Dios            : 10   3rd Qu.:                        1690.7
##  Mirador de Calasanz               :  6   Max.   :                        999310.9
##  (Other)                           :434
##    SHAPELEN            COD_BARRIO              COD_COMUNA            NOM_BARRIO
##  Min.   :  0.029   AUC2   : 259   07     : 361   San Antonio de Prado:      259
##  1st Qu.: 65.461   0903   :  82   16     : 344   BombonÃ¡ No.2       :       82
##  Median : 129.099   0602   :  78   06     : 339   Doce de Octubre No.1:      78
##  Mean   : 242.048   1311   :  71   14     : 333   Belencito          :       71
##  3rd Qu.: 273.097   0608   :  66   13     : 307   Picacho            :       66
##  Max.   :8105.479   0713   :  62   80     : 271   Aures No.1         :       62
##                  (Other):3125   (Other):1788   (Other)            :     3125
```

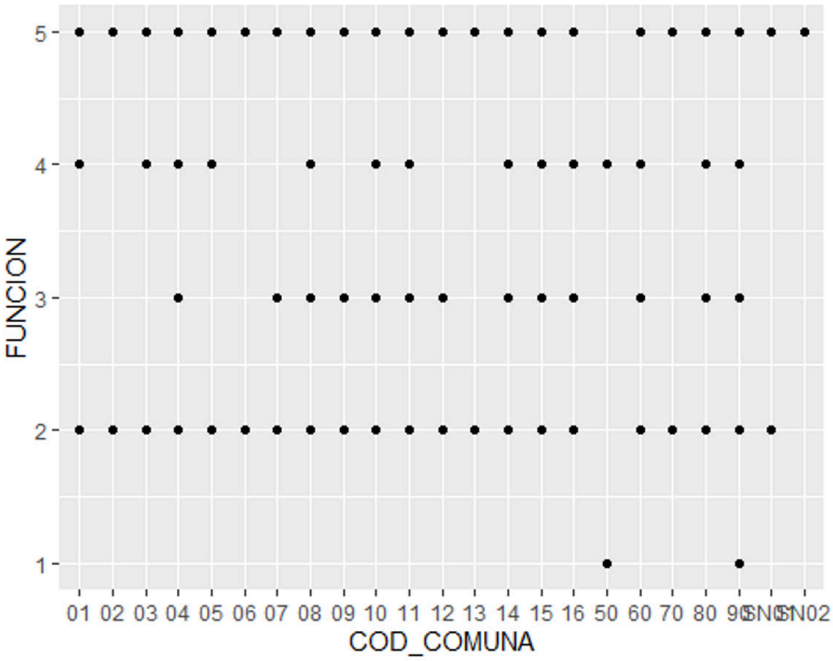
Los diagramas o gráficos de dispersión, también conocidos como nubes de puntos, *scatter plots* o *scatter chart* por su nombre en inglés, representan las observaciones de las variables (generalmente dos, pero también pueden ser tres). El uso principal de un gráfico de dispersión en R es verificar visualmente si existe alguna relación entre ciertas variables numéricas (R-Coder, 2023).

```
ggplot(espacio_publico)+

 geom_point(aes(x= COD_BARRIO, y= FUNCION))
```



► **Figura 11.** Ejemplo de un gráfico de dispersión utilizando el código del barrio y la función del espacio público

```
ggplot(espacio_publico)+

 geom_point(aes(x= COD_COMUNA, y= FUNCION))
```



► **Figura 12.** Ejemplo de un gráfico de dispersión utilizando el número de la comuna y la función del espacio público

```
paradas_colectivos <- read.csv("Paradas_de_Transporte_Público_Colectivo.csv", stringsAs-
Factors = TRUE)

summary(paradas_colectivos)

##      ï..X              Y                    ID_PARADERO                  ID_PARADA
## Min.   :-75.70   Min.   :6.153   Min.   :50001   80001001:  1
## 1st Qu.:-75.60   1st Qu.:6.239   1st Qu.:51542   80001002:  1
## Median :-75.58   Median :6.255   Median :52758   80001003:  1
## Mean   :-75.58   Mean   :6.254   Mean   :52682   80001004:  1
## 3rd Qu.:-75.56   3rd Qu.:6.277   3rd Qu.:53807   80001005:  1
```

```
##  Max.   :-75.52   Max.   :6.351   Max.   :55262   80001006:   1
##                                   (Other) :12713
##    ID_RUTA                NRO_PARADA                DIRECCION
##  Min.   :80001            Min.   : 1.00             CR 64C x CL 67, Medellin:   36
##  1st Qu.:90071            1st Qu.: 10.00            CR 62 x CL 50, Medellin :   31
##  Median :90152            Median : 21.00            CL 44 x CR 52A, Medellin:   30
##  Mean   :89248            Mean   : 24.46            CL 50 64A-01, Medellin  :   30
##  3rd Qu.:90266            3rd Qu.: 35.00            CL 50 x CR 66, Medellin :   30
##  Max.   :90353            Max.   :103.00            CL 67 x DG 64E, Medellin:   29
##                           (Other)                   :12533
##       TIPO_PARADA              RECORRIDO              CODIGO_RUTA
##  Reglamentaria:12719           Destino-Origen:6083   C23i   :  521
##                                Origen-Destino:6636   255A   :  276
##                                103    :  206
##                                95     :  170
##                                107    :  165
##                                24     :  156
##                                (Other):11225
##                        NOMBRE_RUTA     SISTEMA_RUTA
##  Circular                    :  292   8A     :2206
##  Alpujarra - Oriental sentido horario:  121   2A     :1998
```

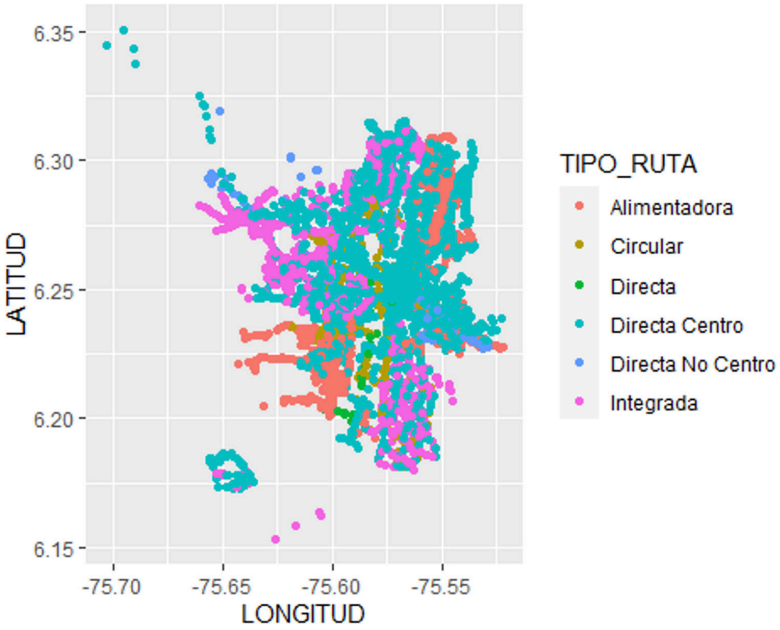
```
##  San Antonio de Prado             :  116            :1785
##  Las Flores-Moravia-Centro        :  103   2B       :1334
##  Doce de Octubre                  :  102   5A       : 880
##  Cerros de Quintalinda            :  100   6H       : 687
##  (Other)                          :11885   (Other):3829
##            TIPO_RUTA                          EMPRESA
##  Alimentadora     :1162          Autobuses Poblado Laureles       :1238
##  Circular         :1021          Transporte MedellÃ¬n Castilla     :1005
##  Directa          :  36          Palenque Robledal                : 892
##  Directa Centro   :7890          RÃ¡pido San CristÃ³bal            : 843
##  Directa No Centro: 405          Tax Maya                         : 767
##  Integrada        :2205          Sistema Alimentador Oriental S.A.S.: 708
##                                  (Other)                          :7266
##     OBJECTID        X_MAGNAMED        Y_MAGNAMED                LONGITUD
##  Min.   :    1     Min.   :820081     Min.   :1172422   Min.          :-75.70
##  1st Qu.: 3340     1st Qu.:831508     1st Qu.:1181855   1st Qu.       :-75.60
##  Median : 6636     Median :834216     Median :1183712   Median        :-75.58
##  Mean   : 6678     Mean   :833581     Mean   :1183611   Mean          :-75.58
##  3rd Qu.: 9924     3rd Qu.:835902     3rd Qu.:1186056   3rd Qu.       :-75.56
##  Max.   :13707     Max.   :840111     Max.   :1194238   Max.          :-75.52
##
```

```
##    LATITUD

## Min.   :6.153

## 1st Qu.:6.239

## Median :6.255

## Mean   :6.254

## 3rd Qu.:6.277

## Max.   :6.351

##
```

En este data set tenemos el dato de longitud y latitud, con esta información se puede hacer un gráfico de puntos para ver cómo lucen en el espacio:

```r
ggplot(paradas_colectivos) +
  geom_point(aes(x=LONGITUD, y=LATITUD, color=TIPO_RUTA))
```

```r
ggplot()+
  geom_sf(data = economia)+
  geom_point(data = paradas_colectivos,(aes(x=LONGITUD, y=LATITUD,
color=TIPO_RUTA)))+
  labs(title="CIUDAD DE MEDELLIN")
```



► **Figura 13.** Ejemplo de un gráfico utilizando los datos de longitud y latitud para identificar los tipos de ruta de transporte público en la ciudad de Medellín

Cuando en una base de datos podemos encontrar información relacionada con la latitud y longitud esto hace referencia a un lugar geográfico. Los Gobiernos proporcionan datos espaciales de libre consulta para diversos procesos sociales y económicos. En plataformas de acceso abierto como Google Maps o redes sociales como Twitter y Facebook también generan datos geoespaciales.

El uso de este tipo de datos permite trabajar con conceptos espaciales como distancia, ubicación, proximidad, vecindad y región. Al incluir estos conceptos se enriquece el análisis de una situación problema (Urdinez y Cruz, 2021).

Esta cartilla pretende ser una motivación para profundizar en el análisis de situaciones problema utilizando el lenguaje de programación R, una herramienta muy versátil que cuenta con un software libre y en constante actualización. Los países tienen a su disposición muchos datos sociales y económicos de libre acceso que si se procesan y analizan correctamente pueden ser un referente importante en la toma de decisiones de la administración de recursos tanto en la empresa pública como en la empresa privada.

Referencias

Analista de datos: ¿cuál es el perfil de estos profesionales? (2019, noviembre 27). *Unir Revista*. https://www.unir.net/ingenieria/revista/analista-de-datos-cual-es-el-perfil-de-estos-profesionales/

Ballari, D. (2018). *Función ggplot()* de ggplot2. https://rpubs.com/dani-ballari/ggplot

Bembibre, V. (2009, febrero). *Definición de histograma*. Definición ABC. https://www.definicionabc.com/tecnologia/histograma.php

Coppola, M. (2022). *JSON para principiantes: qué es, para qué sirve y ejemplos*. https://blog.hubspot.es/website/que-es-json

Fernández, R., y Cotos, T. (2022, enero 11). *Estadística espacial con R*. https://rubenfcasal.github.io/estadistica_espacial/index.html

Instituto Nacional de Estadística. (s. f.). Sobre datos y gráficos [Presentación sin título]. https://www.ine.es/explica/docs/pasos_tipos_graficos.pdf

Introduction to importing data in R. (s. f.). https://campus.datacamp.com/courses/introduction-to-importing-data-in-r/importing-data-from-flat-files-with-utils?ex=3

Lenguaje R, ¿qué es y por qué es tan usado en big data? (2019, noviembre 29). *Unir Revista*. https://www.unir.net/ingenieria/revista/lenguaje-r-big-data/

R-Coder. (2023). *Gráfico de dispersión en R*. https://r-coder.com/grafico-dispersion-r/

Urdinez, F., y Cruz, A. (2021). *AnalizaR Datos Políticos*. https://arcruz0.github.io/libroadp/

Vazquez, A. (2019, mayo 17). *Ciencia de datos para gente sociable*. Health Big Data. https://www.juanbarrios.com/curso-ciencia-de-datos-usando-r/

Wickham, H. (s. f.). *Conflictos entre el tidyverse y otros paquetes*. https://tidyverse-tidyverse-org.translate.goog/reference/tidyverse_conflicts.html?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es-419&_x_tr_pto=sc

Wickham, H., y Grolemund, G. (2019). *R para ciencia de datos*. https://es.r4ds.hadley.nz/index.html