

Prueba Data Engineer Junior

Nombre: Diana Alejandra Bermúdez Fajardo

EXTRACCIÓN:

1. Realicé la extracción en la API del gobierno y convertí los datos de formato JSON en CSV para trabajar de forma más cómoda, utilicé <https://retool.com/utilities>.

The screenshot shows the Retool website's 'Convert JSON to CSV' utility. The page has a header with the Retool logo and navigation links. The main content area is titled 'Upload or paste JSON' and includes a text area for pasting JSON data. To the right of the text area is a table preview showing the resulting CSV data. The table has columns: orden, fecha, a_o, mes, and d_a. The data rows show information for January 1, 2012, including details about a vehicle accident in San Francisco.

Nota: El programa con el que decidí trabajar fue con R studio porque manejo más las librerías y comandos de R que de Python jupyter, aunque son muy similares.

2. Luego en el R realicé la extracción de los datos CSV

The screenshot shows an R Studio script with the following code:

```
9 install.packages("tidyr")
10
11 library(readr)
12 library(readxl)
13 library(janitor)
14 library(dplyr)
15 library(ggplot2)
16 library(tidyr)
17
18
19 #EXTRACCION
20 #1. Leer adecuadamente la base de datos en csv
21
22 Data = read.csv('data.csv', header = TRUE, sep = ",")
23 View(Data)
24
25 names(Data)
26 summary(Data)
27 typeof(Data)
28 str(Data)
29
30 #LIMPIEZA
31 #2. Eliminar columnas redundantes
32
33 Data<-select(Data, -(a_o, mes, d_a, entidad))
34 View(Data)
35
```

The script is executed in the R console, showing the output of the commands. The Environment pane on the right shows the 'Data' object with 1000 observations and 26 variables. The Files pane shows the 'data.csv' file in the 'Downloads' folder.

Se observa que el Data se leyó correctamente:

The screenshot shows the RStudio interface. The top pane displays a data frame with columns: orden, fecha, a_o, mes, d_a, gravedad, peaton, automovil, campero, and car. The right pane shows the 'Data' environment with 1000 observations and 26 variables. The bottom pane shows the console with the following commands and output:

```
R 4.3.2 - C:/Users/diana.bermudez/Downloads/ARCHIVOS_R/
m." "06:30:00 p. m." ...
$ entidad          : chr "AGENTES DTB" "AGENTES DTB" "AGENTES DTB" "AGEN
TES DTB" ...
$ nombrecomuna     : chr "17. MUTIS" "02. NORORIENTAL" "12. CABECERA DEL
LLANO" "03. SAN FRANCISCO" ...
$ propietario_de_veh_culo: chr "Particular" "Empresa" "Particular" "Particula
r" ...
$ diurnio_nocturno  : chr "Diurno" "Diurno" "Diurno" "Nocturno" ...
$ hora_restriccion_moto : chr "No aplica" "No aplica" "No aplica" "No aplica"
```

3. Con ayuda de los comandos puede hacer unos chequeos a la base de datos, como el tipo en que lee las variables, los nombres de las columnas, etc.

The screenshot shows the RStudio interface with the following R code in the script editor:

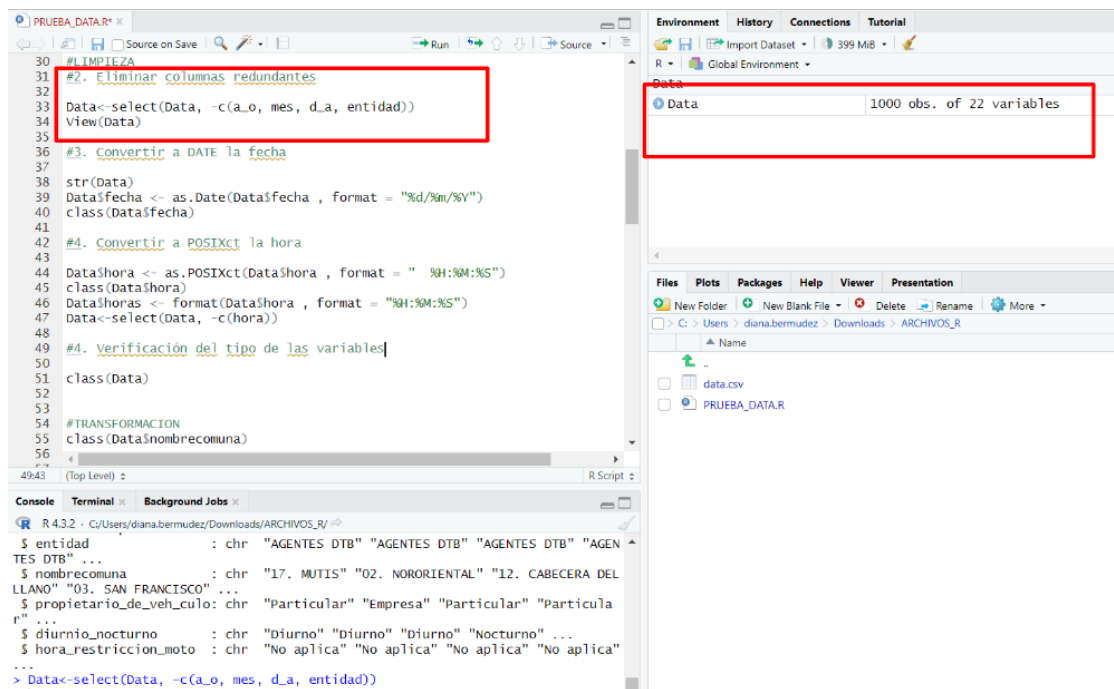
```
9 install.packages("tidyr")
10
11 library(readr)
12 library(readxl)
13 library(janitor)
14 library(dplyr)
15 library(ggplot2)
16 library(tidyr)
17
18
19 #EXTRACCION
20 #1. Leer adecuadamente la base de datos en csv
21
22 Data = read.csv('data.csv', header = TRUE, sep = ",")
23 View(Data)
24
25 names(Data)
26 summary(Data)
27 typeof(Data)
28 str(Data)
29
30 #LIMPIEZA
31 #2. Eliminar columnas redundantes
32
33 Data<-select(Data, -c(a_o, mes, d_a, entidad))
34 View(Data)
35
```

The console output shows the results of the commands:

```
R 4.3.2 - C:/Users/diana.bermudez/Downloads/ARCHIVOS_R/
$ bicicleta        : int 0 0 0 0 0 0 0 0 0 ...
$ otro             : int 0 0 0 0 0 0 0 0 1 0 ...
$ via_1            : chr "CALLE" "VIA MATANZA" "CARRERA" "CARRERA" ...
$ barrio           : chr "Mutis" "Regaderos Norte" "Cabecera del Llano"
"Norte Bajo" ...
$ hora             : chr "12:15:00 p. m." "02:00:00 p. m." "12:00:00 p.
m." "06:30:00 p. m."
```

LIMPIEZA:

4. Se eliminan las columnas que consideré redundantes, a_o, mes, d_a, entidad. También consideré redundante Via_1 porque es muy incompleta la información pero preferí dejarla.



The screenshot shows the RStudio interface. The script editor on the left contains the following code:

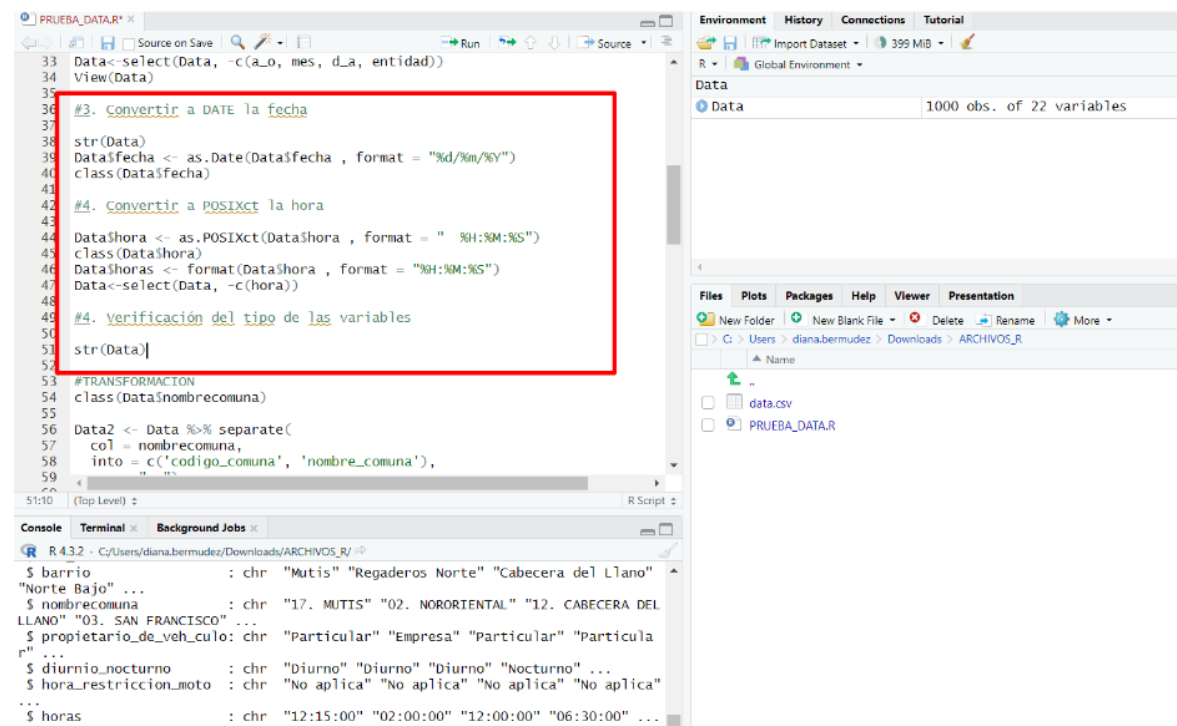
```
#1. Limpieza
#2. Eliminar columnas redundantes
Data<-select(Data, -c(a_o, mes, d_a, entidad))
View(Data)
#3. Convertir a DATE la fecha
str(Data)
Data$fecha <- as.Date(Data$fecha , format = "%d/%m/%Y")
class(Data$fecha)
#4. Convertir a POSIXct la hora
Data$hora <- as.POSIXct(Data$hora , format = " %H:%M:%S")
class(Data$hora)
Data$horas <- format(Data$hora , format = "%H:%M:%S")
Data<-select(Data, -c(hora))
#4. Verificación del tipo de las variables
class(Data)
#TRANSFORMACION
class(Data$nombrecomuna)
```

The console on the bottom left shows the output of the commands:

```
$ entidad : chr "AGENTES DTB" "AGENTES DTB" "AGENTES DTB" "AGEN
TES DTB" ...
$ nombrecomuna : chr "17. MUTIS" "02. NORORIENTAL" "12. CABECERA DEL
LLANO" "03. SAN FRANCISCO" ...
$ propietario_de_veh_culo: chr "Particular" "Empresa" "Particular" "Particula
r" ...
$ diurno_nocturno : chr "Diurno" "Diurno" "Diurno" "Nocturno" ...
$ hora_restriccion_moto : chr "No aplica" "No aplica" "No aplica" "No aplica"
...
> Data<-select(Data, -c(a_o, mes, d_a, entidad))
```

The Environment pane on the right shows a variable named 'Data' with 1000 observations and 22 variables.

5. Con el comando str() pude ver el tipo de variable que tenían, a partir de esta verificación decidí convertir la fecha en DATE y la hora en POSIXct, de resto las demás variables estaban bien su tipo de dato.



The screenshot shows the RStudio interface. The script editor on the left contains the following code:

```
#3. Convertir a DATE la fecha
str(Data)
Data$fecha <- as.Date(Data$fecha , format = "%d/%m/%Y")
class(Data$fecha)
#4. Convertir a POSIXct la hora
Data$hora <- as.POSIXct(Data$hora , format = " %H:%M:%S")
class(Data$hora)
Data$horas <- format(Data$hora , format = "%H:%M:%S")
Data<-select(Data, -c(hora))
#4. Verificación del tipo de las variables
str(Data)
#TRANSFORMACION
class(Data$nombrecomuna)
Data2 <- Data %>% separate(
  col = nombrecomuna,
  into = c('codigo_comuna', 'nombre_comuna'),
  ...
)
```

The console on the bottom left shows the output of the commands:

```
$ barrio : chr "Mutis" "Regaderos Norte" "Cabecera del Llano"
"Norte Bajo" ...
$ nombrecomuna : chr "17. MUTIS" "02. NORORIENTAL" "12. CABECERA DEL
LLANO" "03. SAN FRANCISCO" ...
$ propietario_de_veh_culo: chr "Particular" "Empresa" "Particular" "Particula
r" ...
$ diurno_nocturno : chr "Diurno" "Diurno" "Diurno" "Nocturno" ...
$ hora_restriccion_moto : chr "No aplica" "No aplica" "No aplica" "No aplica"
...
$ horas : chr "12:15:00" "02:00:00" "12:00:00" "06:30:00" ...
```

The Environment pane on the right shows a variable named 'Data' with 1000 observations and 22 variables.

TRANSFORMACIÓN:

6. En la transformación decidí dividir la columna de nombre comuna en dos ("codigo_comuna", "nombre_comuna") para que pueda realizarse una mejor lectura de la información.

The screenshot displays the R Studio environment. The script editor on the left contains the following R code:

```
48  
49 #4. Verificación del tipo de las variables  
50 str(Data)  
51  
52  
53 #TRANSFORMACION  
54  
55 #5. Dividir la columna de nombrecomuna  
56  
57 class(Data$nombrecomuna)  
58  
59 Data2 <- Data %>% separate(  
60   col = nombrecomuna,  
61   into = c('codigo_comuna', 'nombre_comuna'),  
62   sep = ". ")  
63  
64 #6. Renombrar columnas  
65  
66 names(Data2)  
67 Data3 <- Data2 %>% rename( caso = orden, jornada = diurno_nocturno,  
68   tipo_propietario = propietario_de_veh_culo,  
69   restriccion = hora_restriccion_moto,  
70   via = via_1 )  
71  
72 #TRANSFORMACIÓN  
73  
74
```

The Environment pane on the right shows the following data objects:

Object	Size
Data	1000 obs. of 22 variables
Data2	1000 obs. of 23 variables

The Files pane at the bottom shows the project files: data.csv and PRUEBA_DATA.R.

The Console pane at the bottom shows the output of the R code:

```
R 4.3.2 - C:/Users/diana.bermudez/Downloads/ARCHIVOS_R/...  
$ diurno_nocturno : chr "Diurno" "Diurno" "Diurno" "Nocturno" ...  
$ hora_restriccion_moto : chr "No aplica" "No aplica" "No aplica" "No aplica" ...  
$ horas : chr "12:15:00" "02:00:00" "12:00:00" "06:30:00" ...  
> View(Data)  
> class(Data$nombrecomuna)  
[1] "character"  
> Data2 <- Data %>% separate(  
+   col = nombrecomuna,  
+   into = c('codigo_comuna', 'nombre_comuna'),  
+   sep = ". ")
```

7. Renombré algunas columnas para que sean más fácil de interpretar la información.

The image displays two screenshots of the R Studio interface. The top screenshot shows the R script editor with code for data transformation, and the bottom screenshot shows the R console and environment pane.

Top Screenshot: R Script Editor

```

48 #4. Verificación del tipo de las variables
49
50 str(Data)
51
52 #TRANSFORMACION
53
54 #5. Dividir la columna de nombrecomuna
55
56 class(Data$nombrecomuna)
57
58 Data2 <- Data %>% separate(
59   col = nombrecomuna,
60   into = c('codigo_comuna', 'nombre_comuna'),
61   sep = ". ")
62
63 #6. Renombrar columnas
64
65 names(Data2)
66 Data3 <- Data2 %>% rename (caso = orden, jornada = diurno_nocturno,
67   tipo_propietario = propietario_de_veh_culo,
68   restriccion = hora_restriccion_moto,
69   via = via_1 )
70
71 #TRANSFORMACIÓN
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99

```

Bottom Screenshot: R Console and Environment

Console:

```

R 4.3.2 - C:/Users/diana.bermudez/Downloads/ARCHIVOS_R/
> View(Data)
> class(Data$nombrecomuna)
[1] "character"
> Data2 <- Data %>% separate(
+   col = nombrecomuna,
+   into = c('codigo_comuna', 'nombre_comuna'),
+ )

```

Environment:

Object	Size
Data	1000 obs. of 22 variables
Data2	1000 obs. of 23 variables
Data3	1000 obs. of 23 variables

Data3 Preview:

caso	fecha	gravidad	peaton	automovil	campeaero	camioneta	micro	buseta	bus	camion	volqueta	moto	bicicleta	otro	via
1	2012-01-01	Con Heridos	0	1	0	0	0	0	0	0	0	0	0	0	CALLE
2	2012-01-01	Solo Daños	0	1	0	0	1	0	0	0	0	0	0	0	VIA MATANZA
3	2012-01-01	Solo Daños	0	0	0	1	0	0	0	0	0	0	0	0	CARRERA
4	2012-01-01	Solo Daños	0	1	0	1	0	0	0	0	0	0	0	0	CARRERA
5	2012-01-01	Con Heridos	1	0	0	0	0	0	0	0	0	1	0	0	CARRERA
6	2012-01-01	Solo Daños	0	1	0	0	0	0	0	0	0	1	0	0	CALLE
7	2012-01-02	Con Heridos	1	0	0	0	1	0	0	0	0	0	0	0	CARRERA
8	2012-01-02	Solo Daños	0	0	0	2	0	0	0	0	0	0	0	0	TRANSVERSAL METROPOLITANA
9	2012-01-02	Solo Daños	0	1	0	0	0	0	0	0	0	0	0	1	CARRERA
10	2012-01-02	Con Heridos	0	1	0	0	0	0	0	0	0	0	0	0	CARRERA
11	2012-01-02	Solo Daños	0	2	0	0	0	0	0	0	0	0	0	0	CALLE
12	2012-01-02	Solo Daños	0	0	0	0	0	0	0	1	0	1	0	0	AUTOPISTA NORTE
13	2012-01-02	Con Heridos	0	0	0	0	0	0	0	0	0	1	0	0	CALLE
14	2012-01-03	Con Heridos	0	1	0	0	0	0	0	0	0	1	0	0	CARRERA
15	2012-01-03	Con Heridos	0	0	0	0	0	0	0	0	1	1	0	0	CARRERA
16	2012-01-03	Solo Daños	0	0	0	0	1	0	0	1	0	0	0	0	AUTOPISTA NORTE
17	2012-01-03	Solo Daños	0	0	0	0	1	0	0	0	0	1	0	0	CALLE
18	2012-01-03	Solo Daños	0	1	0	1	0	0	0	0	0	0	0	0	AVENIDA Q SECA

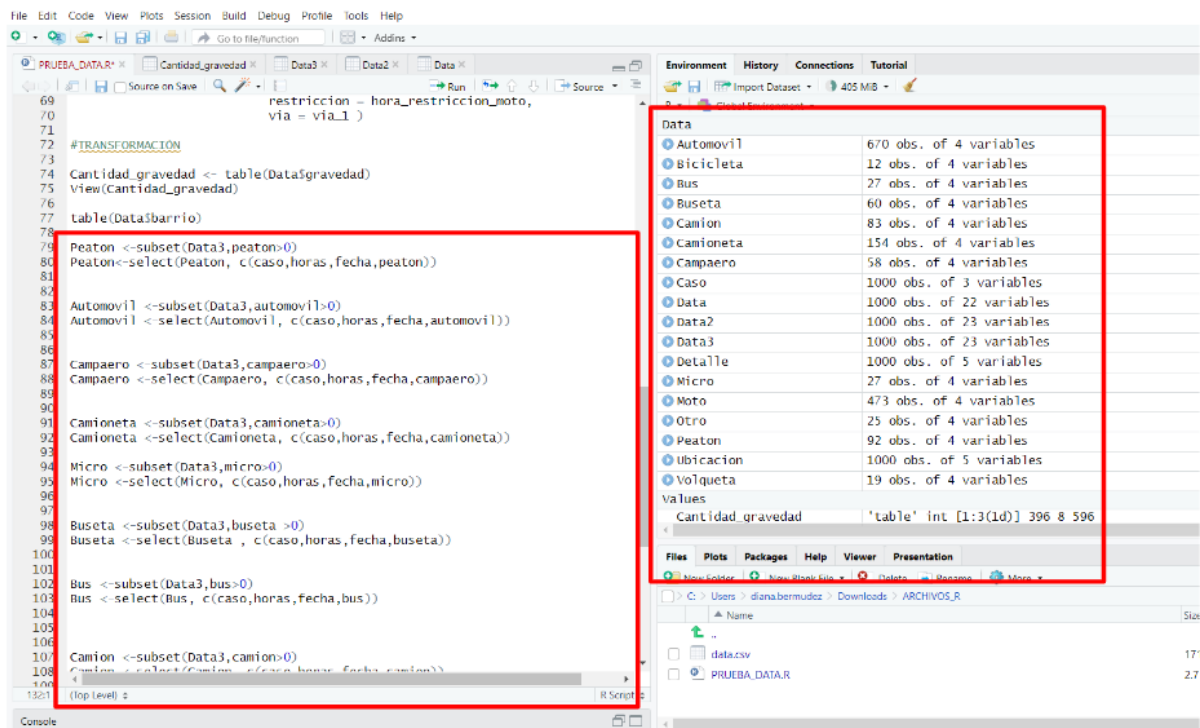
Console (continued):

```

> Data3 <- Data2 %>% rename (caso = orden, jornada = diurno_nocturno,
+   tipo_propietario = propietario_de_veh_culo,
+   restriccion = hora_restriccion_moto,
+   via = via_1 )

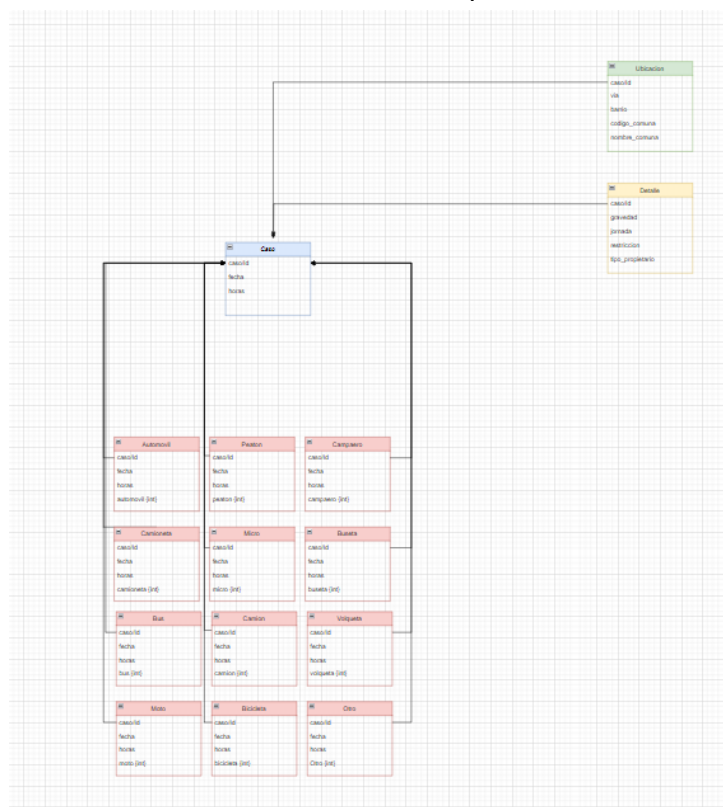
```

8. Finalmente con los comandos `subset()`, `select()` y condicionales, pude dividir la información en varios datas, el campo que seleccione para que se pueda relacionar es con la columna que inicialmente se llamaba “orden”, aunque en el transcurso del procesos renombre a “caso”.



MODELO ENTIDAD RELACIÓN:

- A partir del paso anterior, realicé el diagrama de mis tablas, la tabla principal se llama “Caso”, a continuación se observa el esquema:



FINALIZACIÓN:

“Query SQL”

10. Exporté todas las tablas que creé en R y cargué las tablas en dBeaver donde relacioné tres tablas con un Query por SQL.

The screenshot displays the DBeaver 23.2.4 interface with the following components:

- Top Menu Bar:** Archivo, Editar, Navegar, Buscar, Editor SQL, Base de Datos, Ventana, Ayuda.
- Toolbar:** Includes icons for SQL, Commit, Rollback, Auto, and other database operations.
- Left Panel (Database Explorer):** Shows the 'monitorgit' database structure. The 'Prueba' schema is expanded, showing tables like 'automovil', 'bus', 'camioneta', 'campero', 'caso', 'detalle', 'micro', 'moto', 'otro', 'ubicacion', and 'volqueta'. The 'caso' table is selected.
- SQL Editor:** Contains the following SQL query:


```
select
  c.caso,
  d.gravedad,
  d.restriccion,
  d.tipo_propietario,
  d.jornada,
  u.via,
  u.barrio,
  u.codigo_comuna,
  u.nombre_comuna
from
  "Prueba".caso c
left join "Prueba".ubicacion u
  on c.caso = u.caso
left join "Prueba".detalle d
  on c.caso = d.caso
where
  d.tipo_propietario = 'Particular'
  and d.jornada = 'Diurno'
group by
  c.caso,
  d.gravedad,
  d.restriccion,
  d.tipo_propietario,
  d.jornada,
  u.via,
  u.barrio,
  u.codigo_comuna,
  u.nombre_comuna
order by
  1 asc
```
- Results Panel:** Shows the execution results of the query. The first row is highlighted.

	caso	gravedad	restriccion	tipo_propietario	jornada	via	barrio	codigo_comuna	nombre_comuna
1	3	Solo Daños	No aplica	Particular	Diurno	CARRERA	Cabecera del Llano	12	CABECER
- Bottom Panel:** Includes a 'Project - General' tab and a 'Data Source' section.

blancos
detalle
automovil
bus
camioneta
campaero
micro
moto
otro
ubicacion
volqueta

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23

d.tipo_propietario ,
 d.jornada ,
 u.via ,
 u.barrio ,
 u.codigo_comuna ,
 u.nombre_comuna

caso(+)
1

select c.caso , d.gravedad , d.restriccion , d.tipo
Enter a SQL expression to filter results (use Ctrl+Space)

	123 caso	abc gravedad	abc restriccion	abc tipo_propietario	abc jornada	abc via	abc barrio	abc codigo_comuna
1	1	Con Heridos	No aplica	Particular	Diurno	CALLE	Mutis	17
2	3	Solo Daños	No aplica	Particular	Diurno	CARRERA	Cabecera del Llano	12
3	10	Con Heridos	No aplica	Particular	Diurno	CARRERA	Tejar Norte	01
4	13	Con Heridos	Sin restriccion	Particular	Diurno	CALLE	Fontana	10
5	14	Con Heridos	Sin restriccion	Particular	Diurno	CARRERA	Provenza	10
6	18	Solo Daños	No aplica	Particular	Diurno	AVENIDA Q	Centro	15
7	22	Solo Daños	Sin restriccion	Particular	Diurno	CARRERA	Bolarqui	12
8	23	Solo Daños	No aplica	Particular	Diurno	AVENIDA Q	Los Pinos	13
9	24	Solo Daños	No aplica	Particular	Diurno	CARRERA	Manuela Beltran	11
10	27	Con Heridos	Sin restriccion	Particular	Diurno	CARRERA	Alarcon	03
11	31	Con Heridos	Sin restriccion	Particular	Diurno	CALLE	Centro	15
12	33	Solo Daños	No aplica	Particular	Diurno	CALLE	Alarcon	03
13	34	Con Heridos	Sin restriccion	Particular	Diurno	AVENIDA Q	Nuevo Sotomayor	12
14	35	Con Heridos	No aplica	Particular	Diurno	AVENIDA Q	Alarcon	03
15	36	Solo Daños	No aplica	Particular	Diurno	ANILLO VIA	Rio de Oro I	04
16	37	Con Heridos	No aplica	Particular	Diurno	CARRERA	Antonia Santos Cent	13
17	38	Solo Daños	No aplica	Particular	Diurno	AUTOPISTA	Asturias	09
18	39	Solo Daños	No aplica	Particular	Diurno	CALLE	La Aurora	13
19	41	Con Heridos	Sin restriccion	Particular	Diurno	CARRERA	Cabecera del Llano	12
20	43	Solo Daños	No aplica	Particular	Diurno	CALLE 45 VI	Rio de Oro I	04
21	44	Solo Daños	No aplica	Particular	Diurno	AVENIDA C	Real de Minas	07
22	47	Solo Daños	No aplica	Particular	Diurno	CARRERA	Bolarqui	12
23	49	Con Heridos	Sin restriccion	Particular	Diurno	CALLE	La Ceiba	06