Universidad del Valle de Guatemala
Data Science 1 - Sección 10
Christopher Kevin Sandoval García 13660
María Fernanda Estrada Cornejo 14198
Luis Estuardo Delgado Ordoñez 1
Estuardo Díaz 1
08 de octubre del 2020

Laboratorio 8

Limpieza y procesamiento de datos

Los siguientes pasos se aplicaron para la limpieza de datos:

1. Eliminar la columna extra que desordena las columnas, para que se encuentren en el orden correcto los tags y los valores.
2. Verificar que los nombres de las columnas coincidan entre archivos.
3. Convertir todas las columnas a mayúsculas
4. Eliminar caracteres especiales, como #, @, comillas o apóstrofes (de cualquier tipo), etc.
5. Verificar que en las columnas de texto solo hubiera texto.
6. Verificar que en las columnas numéricas solo hubieran números.
7. Verificar que los datos hayan sido ingresados correctamente (ej. cambiar de 3015 a 2015)

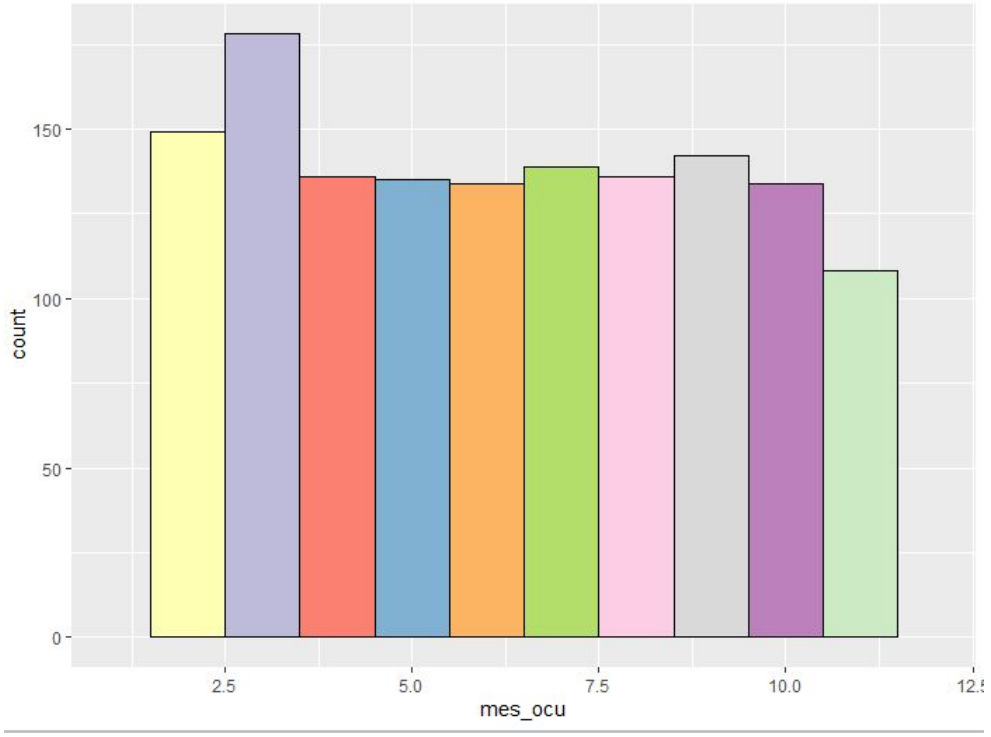
La limpieza de datos se realizó en R, por lo que los módulos utilizados fueron tools, lubridate y stringr. Por otro lado, se dejaron afuera columnas no significativas. Por ejemplo, al trabajar únicamente con motos, la cantidad de asientos siempre era 2, el tonelaje siempre era 0, puertas siempre en 0, etc. Estos tipos de datos se descartaron del análisis.

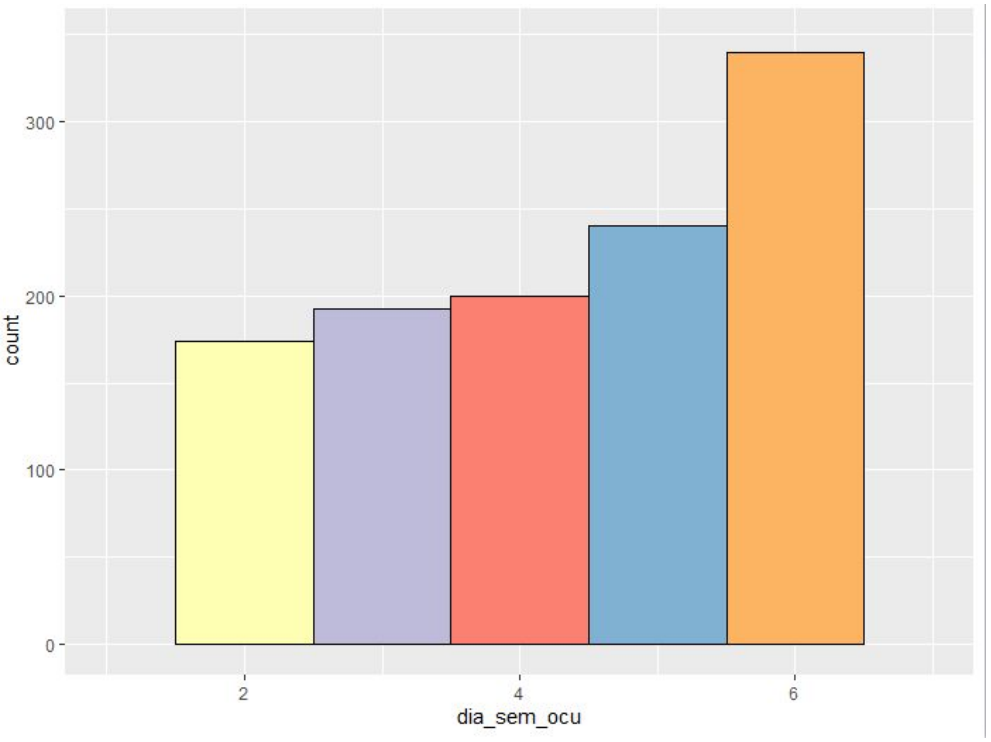
Análisis exploratorio accidentes

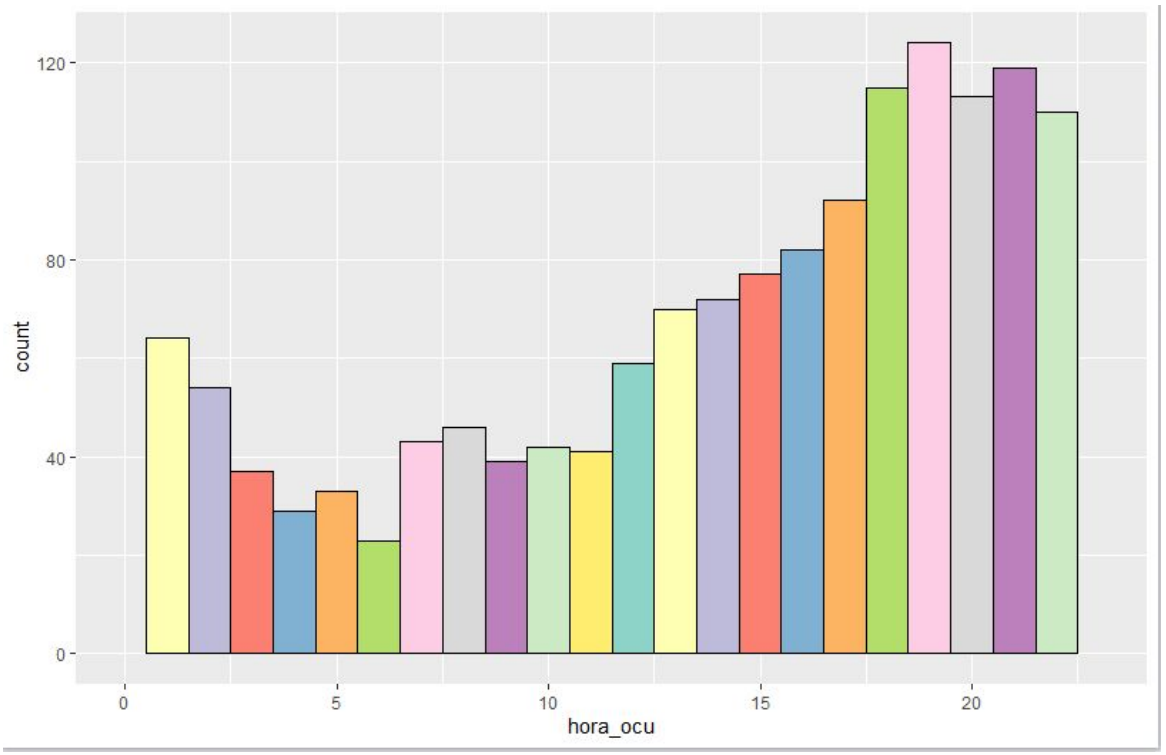
Resumen variables cuantitativas

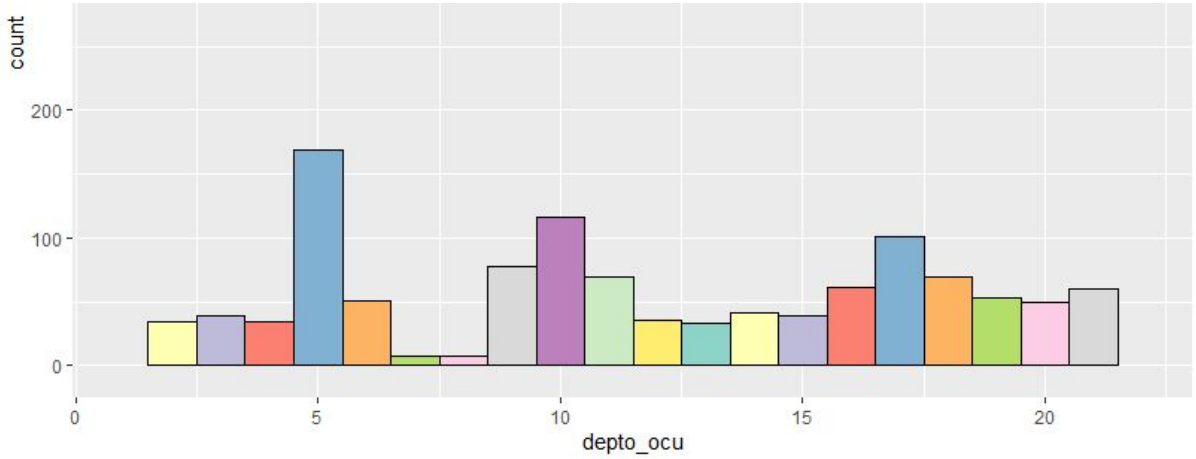
```
    dia_ocu          mes_ocu         dia_sem_ocu       hora_ocu        depto_ocu        sexo_pil    
 Min.   : 1.00   Min.   : 1.000   Min.   :1.000   Min.   : 0.00   Min.   : 1.000   Min.   :1.000  
 1st Qu.: 8.00   1st Qu.: 3.000   1st Qu.:3.000   1st Qu.: 8.00   1st Qu.: 1.000   1st Qu.:1.000  
 Median :16.00   Median : 6.000   Median :5.000   Median :16.00   Median : 9.000   Median :1.000  
 Mean   :15.37   Mean   : 6.337   Mean   :4.478   Mean   :13.69   Mean   : 8.779   Mean   :1.042  
 3rd Qu.:23.00   3rd Qu.: 9.000   3rd Qu.:6.000   3rd Qu.:20.00   3rd Qu.:16.000   3rd Qu.:1.000  
 Max.   :31.00   Max.   :12.000   Max.   :7.000   Max.   :23.00   Max.   :22.000   Max.   :2.000  
    edad_pil       estado_pil       tipo_veh   color_veh        causa_acc     
 Min.   :15.00   Min.   :1.000   Min.   :4   Min.   : 1.000   Min.   :1.000  
 1st Qu.:22.00   1st Qu.:1.000   1st Qu.:4   1st Qu.: 1.000   1st Qu.:1.000  
 Median :25.00   Median :1.000   Median :4   Median : 3.000   Median :1.000  
 Mean   :28.12   Mean   :1.295   Mean   :4   Mean   : 3.228   Mean   :1.862  
 3rd Qu.:32.00   3rd Qu.:2.000   3rd Qu.:4   3rd Qu.: 5.000   3rd Qu.:2.000  
 Max.   :72.00   Max.   :2.000   Max.   :4   Max.   :16.000   Max.   :5.000  
```

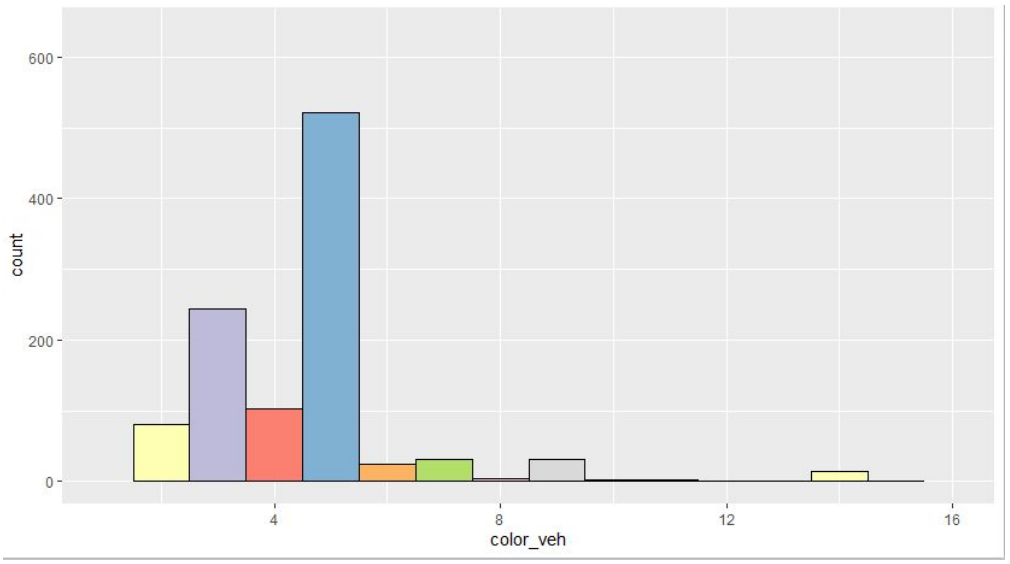
Histogramas variables cuantitativas

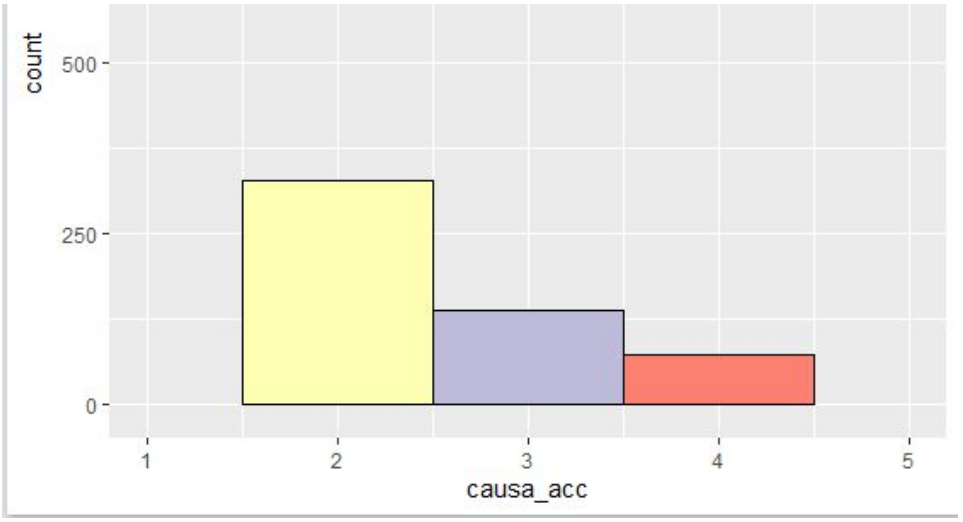




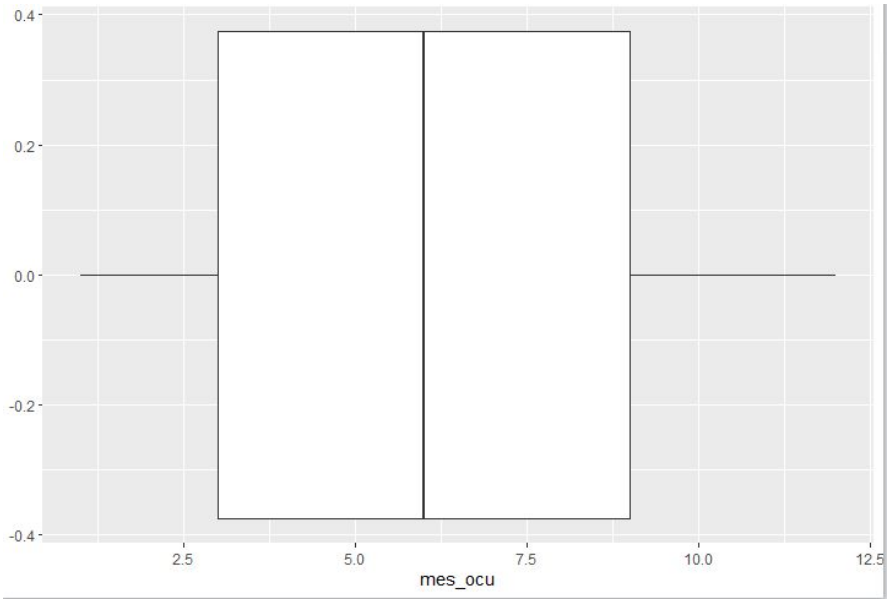

120
80
40
0
count
0
5
10
15
20
hora_ocu

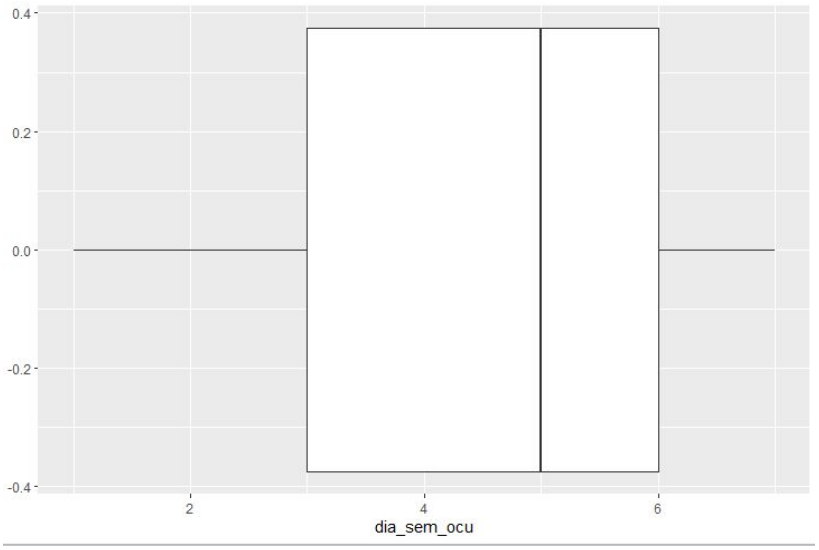

count
200
100
0
0
5
10
15
20
depto_ocu

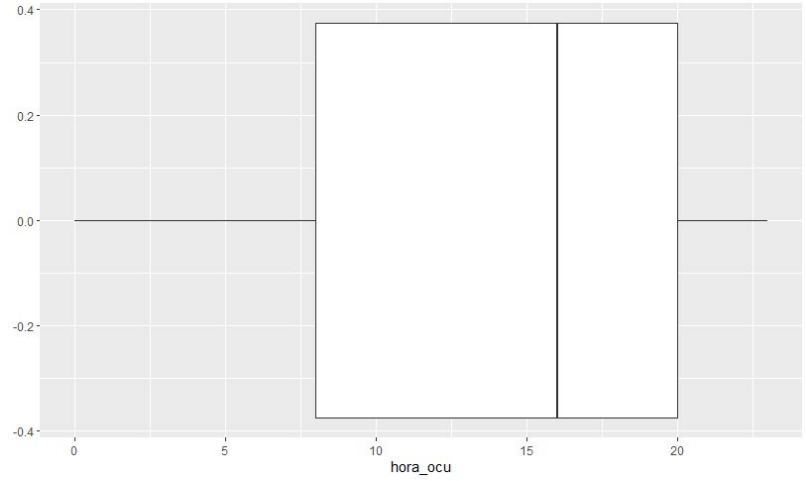




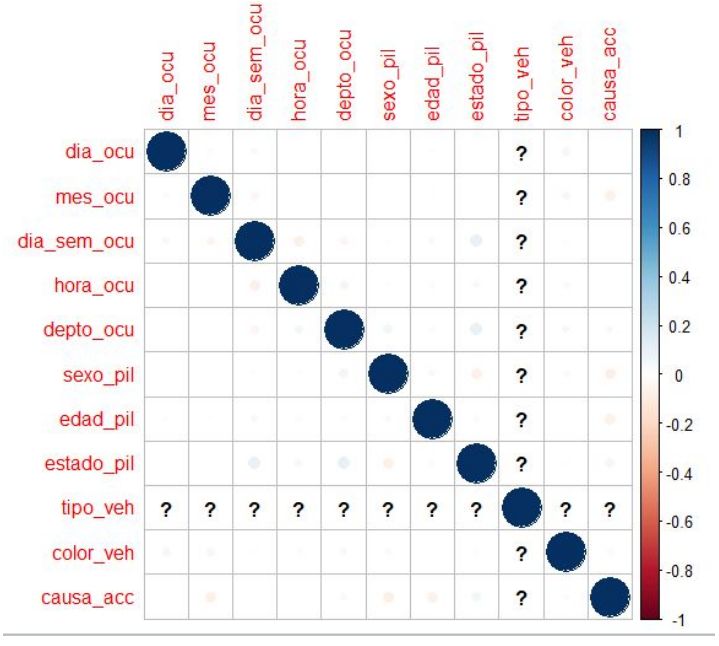
Caja y bigotes variables cuantitativas







Correlación variables cuantitativas



Tablas de frecuencias variables cualitativas

Mes de ocurrencia

	Var1	Freq
3	3	178
12	12	156
2	2	149
1	1	148
9	9	142
7	7	139
4	4	136
8	8	136
5	5	135
6	6	134
10	10	134
11	11	108

Dia de la semana de ocurrencia

	Var1	Freq
7	7	347
6	6	339
5	5	240
1	1	203
4	4	200
3	3	192
2	2	174

Hora de ocurrencia

	Var1	Freq
20	19	124
22	21	119
19	18	115
21	20	113
23	22	110
1	0	109
24	23	102
18	17	92
17	16	82
16	15	77
15	14	72
14	13	70
2	1	64
13	12	59
3	2	54
9	8	46
8	7	43
11	10	42
12	11	41
10	9	39
4	3	37
6	5	33
5	4	29
7	6	23

Sexo del piloto

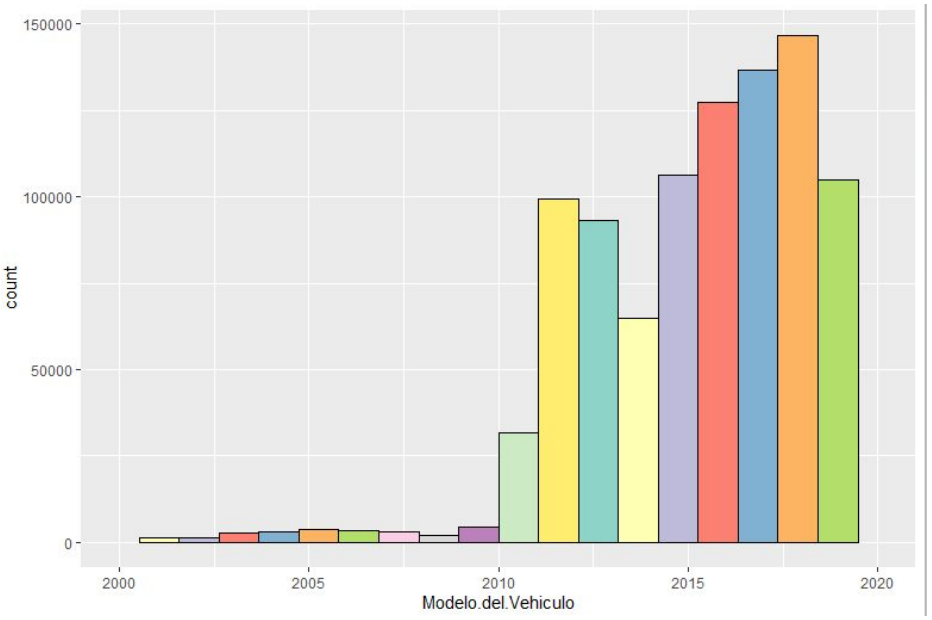
	Var1	Freq
1	1	1624
2	2	71

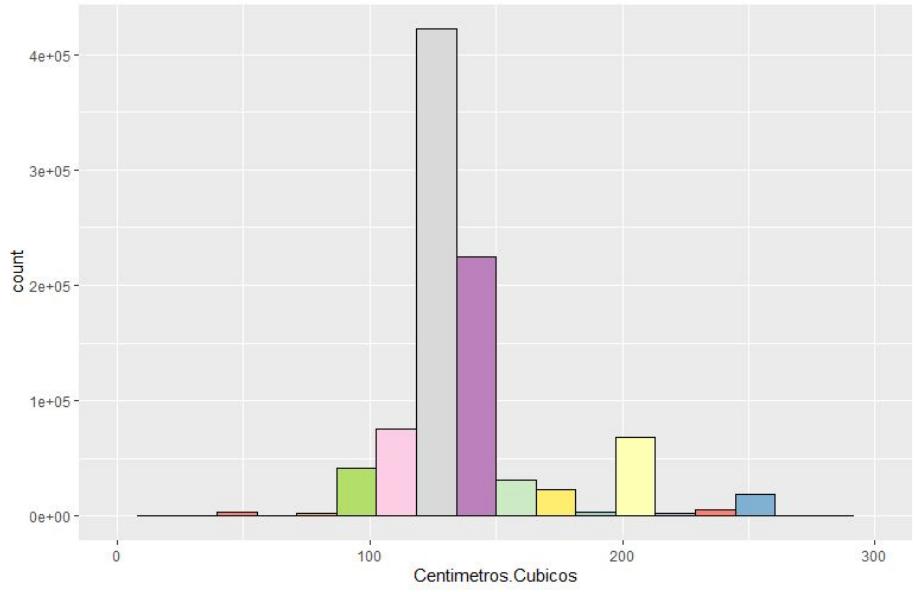
Análisis exploratorio importaciones

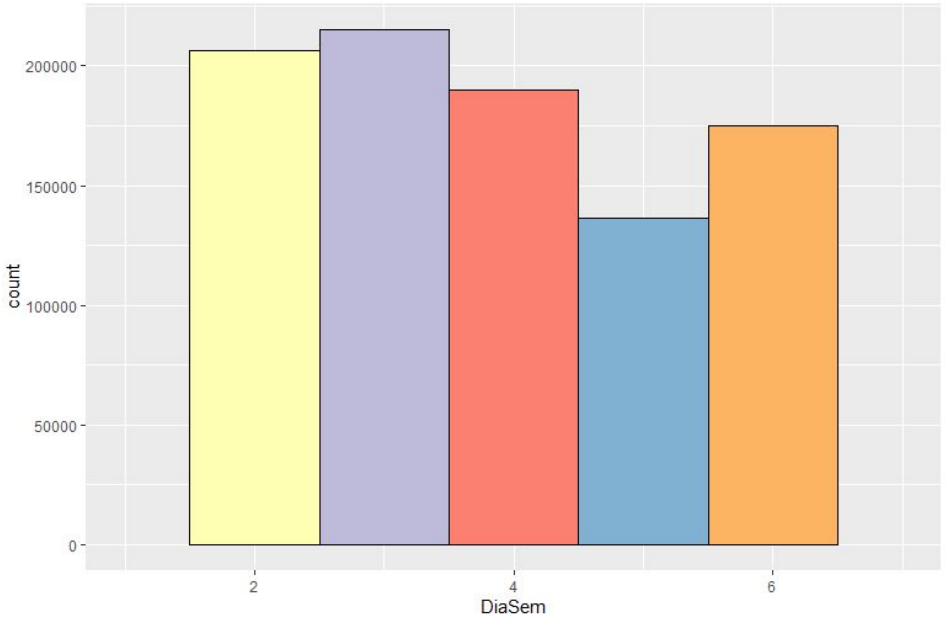
Resumen variables cuantitativas

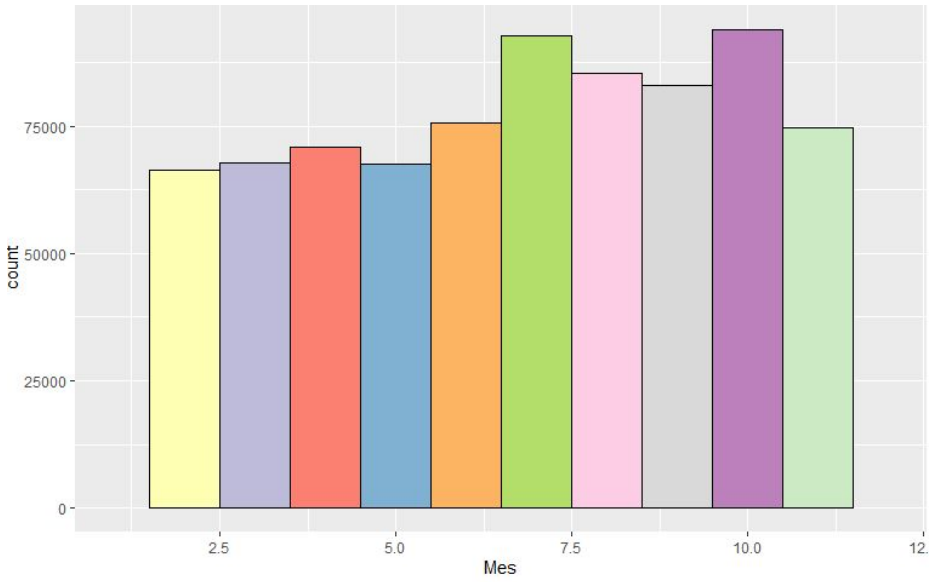
```
Modelo.del.Vehiculo Centimetros.Cubicos   Valor.CIF
Min.   :1900        Min.   :   0.0      Min.   :    467
1st Qu.:2013        1st Qu.: 125.0      1st Qu.: 315627
Median :2016        Median : 125.0      Median : 450144
Mean   :2015        Mean   : 149.1      Mean   : 739615
3rd Qu.:2018        3rd Qu.: 150.0      3rd Qu.: 808493
Max.   :2020        Max.   :6000.0      Max.   :7257974
   Impuesto           Anio           Mes              Dia
Min.   :    56   Min.   :2011   Min.   : 1.000   Min.   : 1.00
1st Qu.: 40526   1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00
Median : 58696   Median :2015   Median : 7.000   Median :15.00
Mean   : 97065   Mean   :2015   Mean   : 6.648   Mean   :15.23
3rd Qu.:109668   3rd Qu.:2017   3rd Qu.:10.000   3rd Qu.:23.00
Max.   :870957   Max.   :2019   Max.   :12.000   Max.   :31.00
    DiaSem
Min.   :1.000
1st Qu.:3.000
Median :4.000
Mean   :3.892
3rd Qu.:5.000
Max.   :7.000
```

Histogramas variables cuantitativas

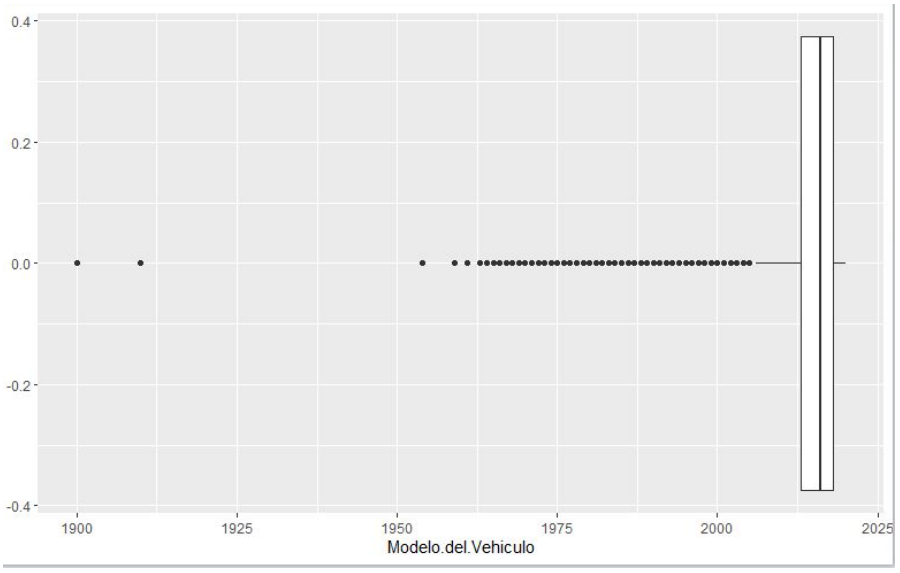


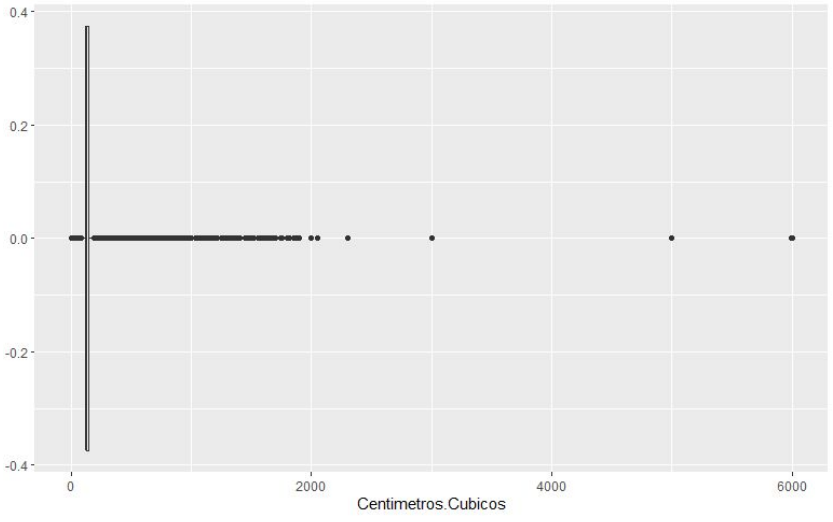

4e+05
3e+05
2e+05
1e+05
0e+00
count
0
100
200
300
Centimetros.Cubicos

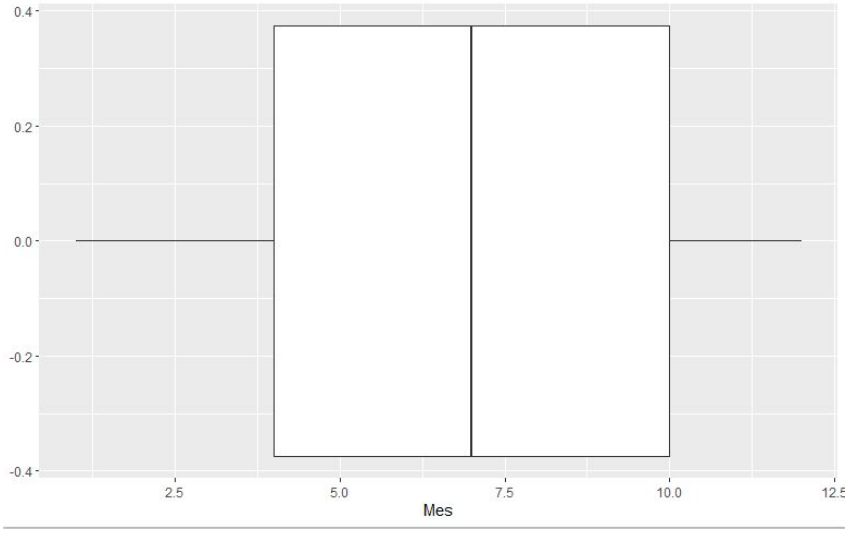

200000
150000
100000
50000
0
count
2
4
6
DiaSem



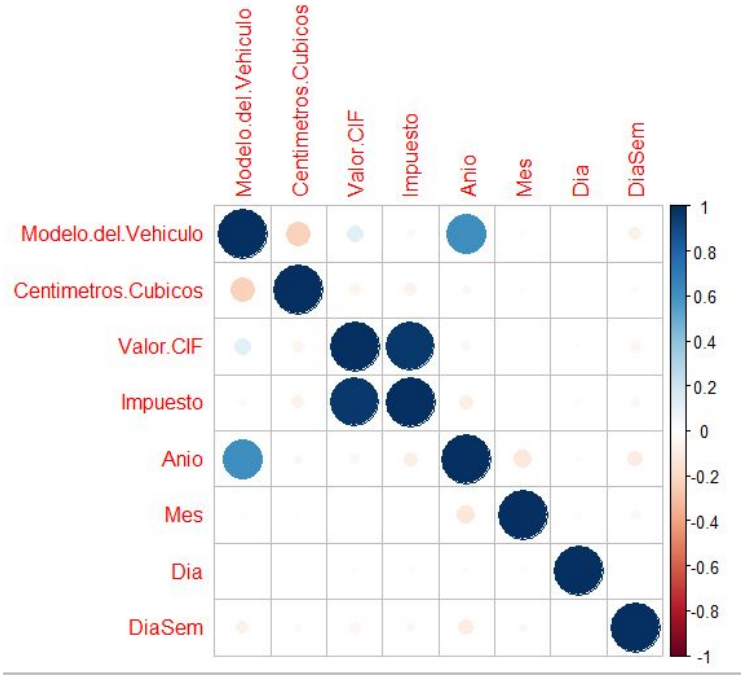
Caja y bigotes variables cuantitativas







Correlación variables cuantitativas



Tablas de frecuencias variables cualitativas

País de proveniencia

	Var1	Freq
12	CHINA	690897
28	INDIA	197620
32	JAPON	20527
9	BRASIL	12084
47	TAIWAN	6280
21	ESTADOS UNIDOS	4865
29	INDONESIA	1804
1	ALEMANIA REP. FED.	1458
46	TAILANDIA	1117
31	ITALIA	1001

Aduana de ingreso

	Var1	Freq
20	PUERTO QUETZAL	803894
5	CENTRAL DE GUATEMALA	61773
8	EXPRESS AEREO	25386
6	EL CARMEN	18847
22	SANTO TOMAS DE CASTILLA	8889
23	TECUN UMAN	8858
18	PEDRO DE ALVARADO	5273
19	PUERTO BARRIOS	3857
1	ADUANA INTEGRADA AGUA CALIENTE	1478
21	SAN CRISTOBAL	1382

Marca más importada

	Var1	Freq
325	SUZUKI	196001
148	HONDA	174507
160	ITALIKA	128453
126	FREEDOM	81086
29	BAJAJ	73196
387	YAMAHA	57999
296	SERPENTO	49580
239	MOVESA	27520
143	HERO	26437
21	ASIA HERO	17130

Decisiones de diseño

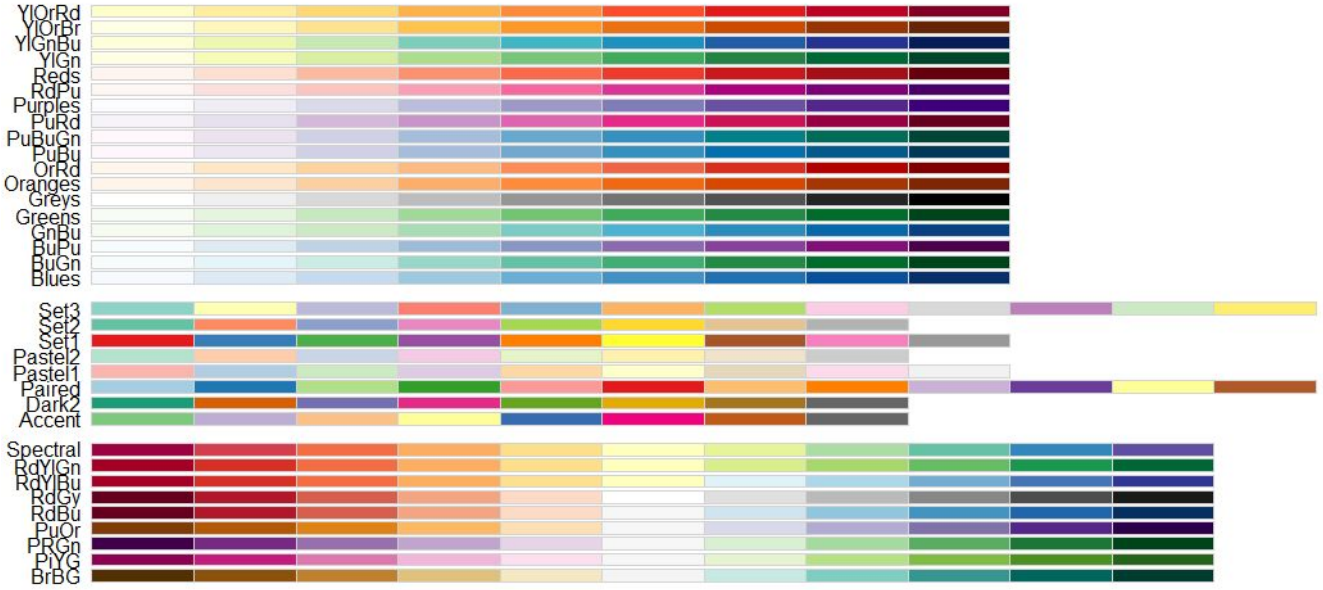
Los gráficos que se colocaron en la infografía fueron los más representativos de los data sets y los que más variables tuvieran. Por ejemplo, en cuanto a importación, se incluyeron gráficos de modelos, centímetros cúbicos, color, etc. En cuanto al de accidentes, se incluyeron los días de ocurrencia, el mes, día de la semana, sexo, etc. Esto se hizo para no aburrir al lector y que observara los datos más relevantes.

No se incluyeron los diagramas de caja y bigotes porque son difíciles de entender para el público general y, como los datos no eran normales, no aportaba mayor información.

Los datos de las tablas de frecuencias se presentaron como texto y no como la imagen colocada en este documento.

La paleta de colores que se decidió utilizar fue una suave, pero con ciertos colores fuertes o llamativos. De esta forma, no se cansa la vista al leer la infografía, pero no se pierde la atención con colores aburridos. Además, al ser datos serios y que provienen de entidades gubernamentales, no se podían usar colores muy diversos.

La paleta de colores seleccionada fue “Set3” de la libreria “RColorBrewer” de R. A continuación, un ejemplo



El orden se presentó por el data set analizado. Primero se presentaron los resultados del análisis exploratorio de importaciones y luego el de accidentes. De esta forma, el dataset de importaciones coloca un precedente al de accidentes.

En general, se intentó utilizar la menor cantidad de texto posible, resaltando solamente la palabra más importante de otro color para llamar más la atención.