

Trabajo de Visualización y Análisis Gráfico con ggplot.

Máster en Data Science y Big Data - Universidad de Sevilla, 2017.

Jerónimo Carranza Carranza

28 de octubre de 2017

Índice

1. Introducción. Descripción del conjunto de datos.	3
1.1. Datos estáticos	3
1.2. Datos dinámicos	3
2. Localización de estaciones	4
3. Datos faltantes	7
4. Datos anómalos	11
5. Análisis de datos válidos globales	14
5.1. Análisis según días de la semana	17
5.2. Análisis según hora del día	20
5.3. Análisis según hora del día y día de la semana	24
6. Análisis de datos válidos por estaciones	25
6.1. Análisis de correlación entre estaciones	25
6.2. Clasificación de las estaciones	30

Índice de cuadros

1. Datos estáticos	3
2. Datos dinámicos	3
3. Resumen de datos anómalos	11
5. Bicis circulantes por día de la semana. Estadística básica.	18
6. Bicis circulantes por hora del día. Estadística básica.	21

Índice de figuras

1. Localización de estaciones SEVICI	5
2. Localización de estaciones SEVICI con Identificadores	6
3. Datos faltantes. Huecos Globales por Fecha y Hora	8
4. Datos faltantes. Huecos Globales por Fecha	9
5. Datos faltantes. Huecos Globales por Hora	9
6. Datos faltantes. Huecos Globales por Mes	10
7. Datos faltantes. Huecos Globales por Día de la Semana	10
8. Datos anómalos. Estacionamientos disponibles > Est. Operativos	12
9. Datos anómalos. Bicicletas disponibles > Est. Operativos	12
10. Datos anómalos. Estacionamientos + Bicicletas disponibles > Est. Operativos	13
11. Datos anómalos. Estacionamientos + Bicicletas disponibles < Est. Operativos	13
12. Datos válidos globales. Distribución de Estacionamientos y Bicis disponibles.	14
13. Datos válidos globales. Número de estaciones disponibles.	14

14. Datos válidos globales. Estacionamientos disponibles.	15
15. Datos válidos globales. Bicis disponibles.	15
16. Datos válidos globales. Diferencia Bicis circulantes y Bicis disponibles.	16
17. Datos válidos globales. Bicis circulantes según día de la semana.	17
18. Datos válidos globales. Bicis disponibles según día de la semana.	18
19. Datos válidos globales. Bicis circulantes por día de la semana. Estadística básica. Mediana, Máximo y Mínimo.	19
20. Datos válidos globales. Bicis circulantes por día de la semana. Media +/- 2 · Desviación	19
21. Datos válidos globales. Bicis circulantes según hora del día.	20
22. Datos válidos globales. Bicis disponibles según hora del día.	21
23. Datos válidos globales. Bicis circulantes por hora del día. Estadística básica. Mediana, Máximo y Mínimo	22
24. Datos válidos globales. Bicis circulantes por hora del día. Estadística básica. Media +/- 2 · Desviación	23
25. Datos válidos globales. Bicis circulantes según hora del día y día de la semana.	24
26. Datos válidos globales. Bicis disponibles según hora del día y día de la semana.	25
27. Datos válidos estaciones. Matriz de correlación ($ corr >0.5$) entre estaciones.	26
28. Datos válidos estaciones. Grafo espacial de correlaciones $ corr > 0.5$	27
29. Datos válidos estaciones. Suma de correlaciones positivas por estación ($corr>0.5$).	28
30. Datos válidos estaciones. Suma de correlaciones negativas por estación ($corr<-0.5$).	29
31. Datos válidos estaciones. Dendrograma de estaciones basado en correlación.	31
32. Datos válidos estaciones. Clasificación de estaciones.	32
33. Datos estaciones. Estacionamientos disponibles por clase de estación, hora del día y día de la semana.	33
34. Datos estaciones. Estacionamientos disponibles por hora del día y día de la semana. Patrones por clase de estación.	35

1. Introducción. Descripción del conjunto de datos.

Los datos que se van a utilizar en este trabajo corresponden a los datos ofrecidos por la empresa JCDecaux en su página web para las ciudades (27) en las que opera los servicios de bicicletas compartidas, y que han sido recopilados por la Universidad de Huelva, durante un año, mediante la captura de datos instantáneos ofrecidos en el servicio web de JCDecaux. Concretamente nos centramos en los datos de la ciudad de Sevilla. Se dispone de dos tipos de datos:

1.1. Datos estáticos

Los denominados datos estáticos hacen referencia a las características de las estaciones. Contiene los siguientes datos para un total de 260 estaciones:

Cuadro 1: Datos estáticos

Campo:	Descripción:
Number	Número de la estación
Name	Nombre de la estación
Address	Dirección
Latitude	Latitud (grados WGS84)
Longitude	Longitud (grados WGS84)

1.2. Datos dinámicos

Los datos dinámicos hacen referencia a la disponibilidad y uso del servicio para cada una de las estaciones incluyendo la siguiente información:

Cuadro 2: Datos dinámicos

Campo:	Descripción:
id	Id registro autonumérico
status	Estado de la estación; OPEN o CLOSED
contract	Contrato, en nuestro caso; Seville
num	Número de la estación
last_update	Momento de última actualización
add_date	Fecha-Hora en fracciones de 5 minutos
stands	Número de estacionamientos operativos en la estación
availablestands	Número de estacionamientos disponibles
availablebikes	Número de bicicletas operativas y disponibles

El número de estacionamientos operativos para cada estación se ha comprobado que es constante, por lo que se trata entre los datos estáticos.

Los datos originales, descritos hasta aquí, se han reorganizado de diversas formas para los diversos pretratamientos y tratamientos realizados. Salvo evidencia al respecto se describirá la organización concreta de los datos utilizados para cada uno de los gráficos o grupos de gráficos que siguen.

2. Localización de estaciones

```
## # A tibble: 6 x 5
##   Number           Name
##   <int>          <chr>
## 1    126 126_AVENIDA REINA MERCEDES
## 2     73 073_PLAZA SAN AGUSTIN
## 3     96 096_CALLE BETIS
## 4    256 256_MIGUEL MONTORO
## 5     82 082_CALLE LUIS MONTOTO
## 6   226 226_AVENIDA DOCTOR EMILIO LEMOS
## # ... with 3 more variables: Address <chr>, Latitude <dbl>,
## #   Longitude <dbl>
## [1] "SpatialPointsDataFrame"
## attr(,"package")
## [1] "sp"
## [1] "sf"        "tbl_df"    "tbl"       "data.frame"
## Simple feature collection with 6 features and 3 fields
## geometry type:  POINT
## dimension:      XY
## bbox:           xmin: -5.999911 ymin: 37.35816 xmax: -5.910091 ymax: 37.40286
## epsg (SRID):    4326
## proj4string:   +proj=longlat +ellps=WGS84 +towgs84=0,0,0,0,0,0,0 +no_defs
## # A tibble: 6 x 4
##   Number           Name
##   <int>          <chr>
## 1    126 126_AVENIDA REINA MERCEDES
## 2     73 073_PLAZA SAN AGUSTIN
## 3     96 096_CALLE BETIS
## 4    256 256_MIGUEL MONTORO
## 5     82 082_CALLE LUIS MONTOTO
## 6   226 226_AVENIDA DOCTOR EMILIO LEMOS
## # ... with 2 more variables: Address <chr>, geometry <simple_feature>
##       xmin     ymin     xmax     ymax
## -6.012005 37.320418 -5.908922 37.424758
```



Figura 1: Localización de estaciones SEVICI

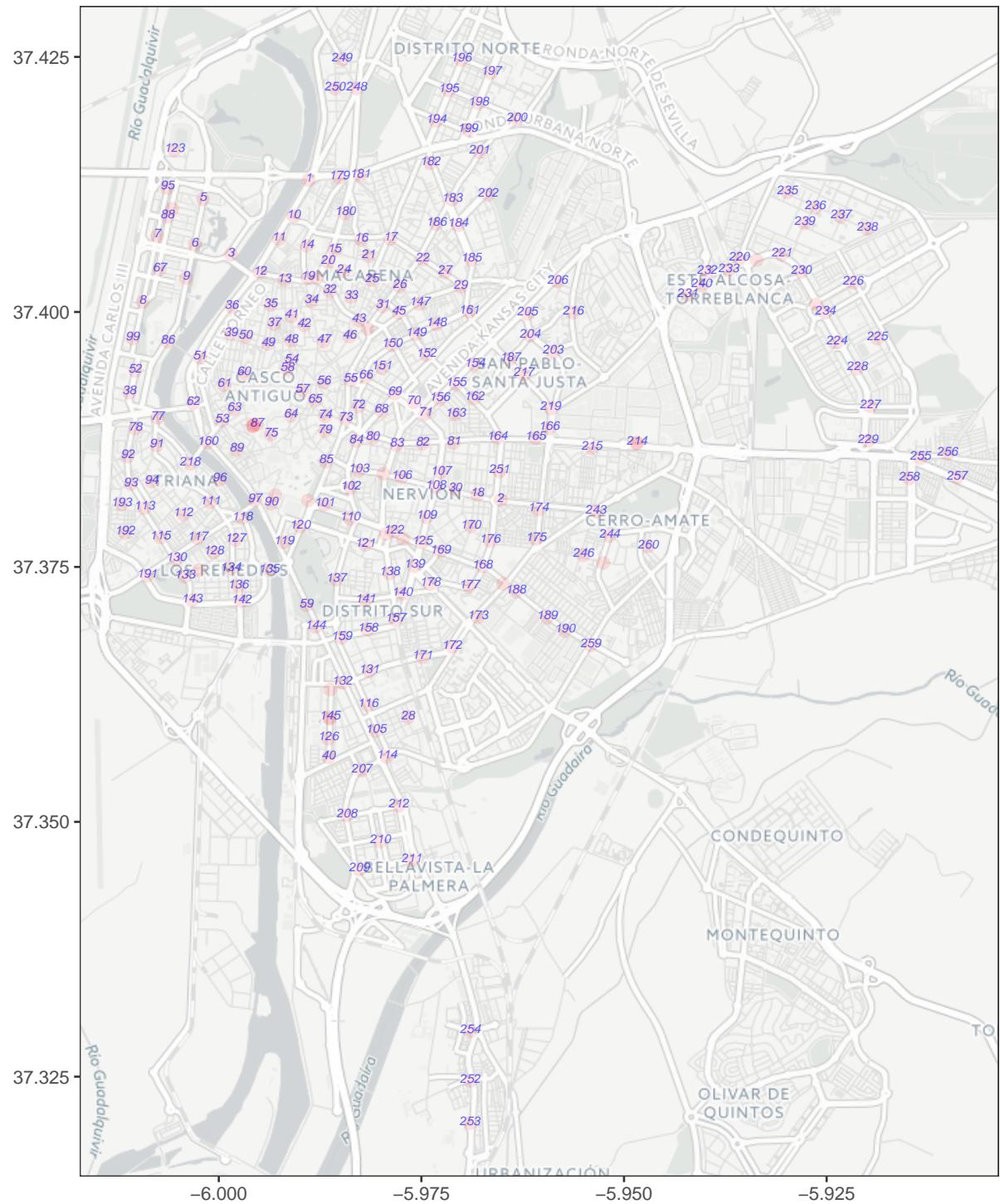


Figura 2: Localización de estaciones SEVICI con Identificadores

3. Datos faltantes

```
## # A tibble: 5 x 10
##       p5min hueco     s1     b1     s2     b2     s3     b3     s4
##   <dttm> <int> <int> <int> <int> <int> <int> <int> <int>
## 1 2015-12-01 00:00:00     0    19     1    17     1    38     0    39
## 2 2015-12-01 00:05:00     0    19     1    17     1    38     0    39
## 3 2015-12-01 00:10:00     0    19     1    17     1    38     0    39
## 4 2015-12-01 00:15:00     0    19     1    17     1    38     0    39
## 5 2015-12-01 00:20:00     0    19     1    17     1    38     0    39
## # ... with 1 more variables: b4 <int>

## # A tibble: 2 x 2
##   hueco     n
##   <int> <int>
## 1     0 103893
## 2     1    1239
```

La variable *hueco* indica si es un hueco global (1) o no (0), es decir, si no existe ningún dato para ninguna estación en ese momento (p5min)(1) o existe al menos una con datos (0).

```
## # A tibble: 6 x 8
##       p5min hueco      DIA    HORAM    HORA    MES  DSEM  DSEMN
##   <dttm> <int> <date> <dbl> <int> <chr> <chr> <dbl>
## 1 2015-12-01 00:00:00     0 2015-12-01 0.00000000     0    12  mar     3
## 2 2015-12-01 00:05:00     0 2015-12-01 0.08333333     0    12  mar     3
## 3 2015-12-01 00:10:00     0 2015-12-01 0.16666667     0    12  mar     3
## 4 2015-12-01 00:15:00     0 2015-12-01 0.25000000     0    12  mar     3
## 5 2015-12-01 00:20:00     0 2015-12-01 0.33333333     0    12  mar     3
## 6 2015-12-01 00:25:00     0 2015-12-01 0.41666667     0    12  mar     3
```

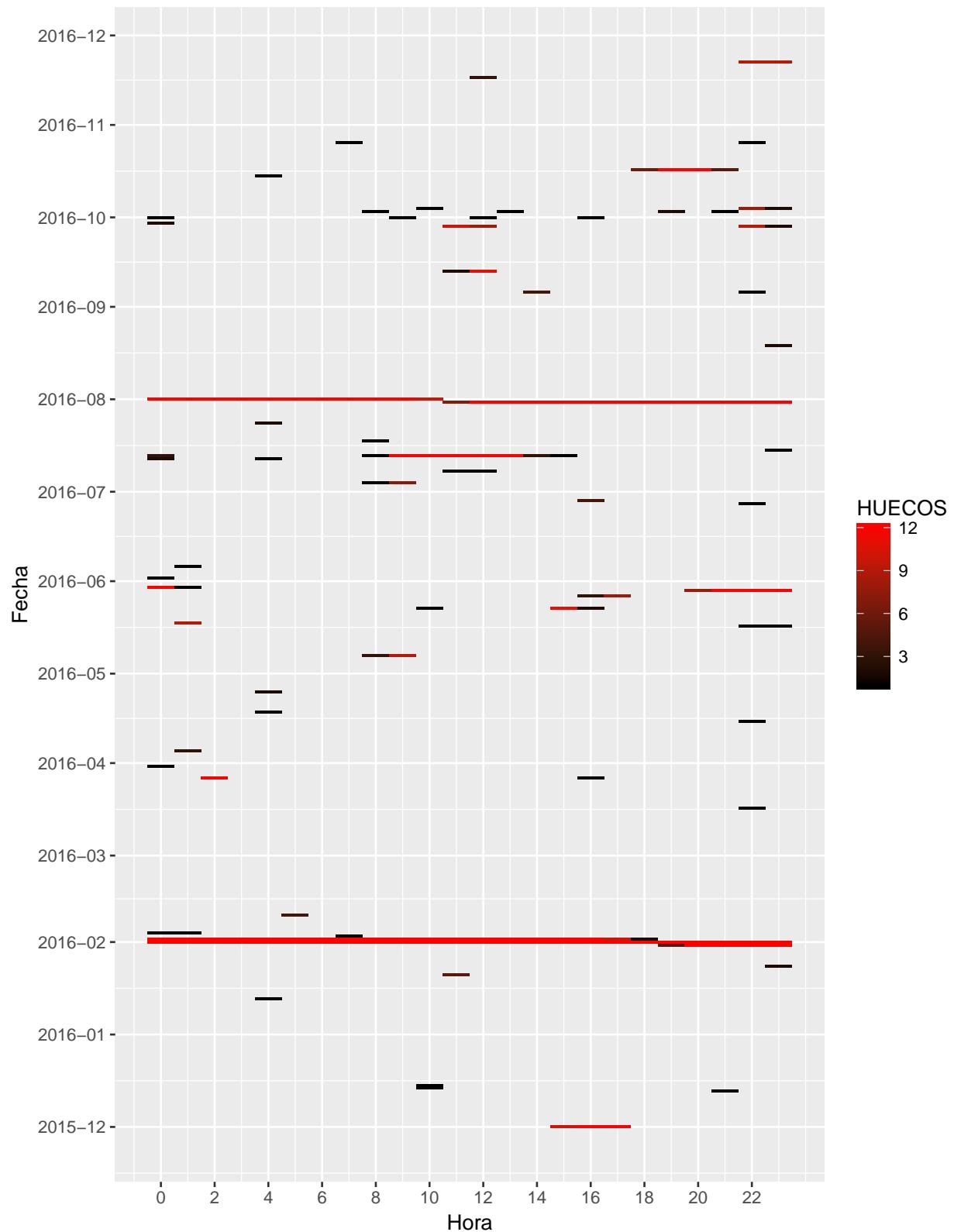


Figura 3: Datos faltantes. Huecos Globales por Fecha y Hora

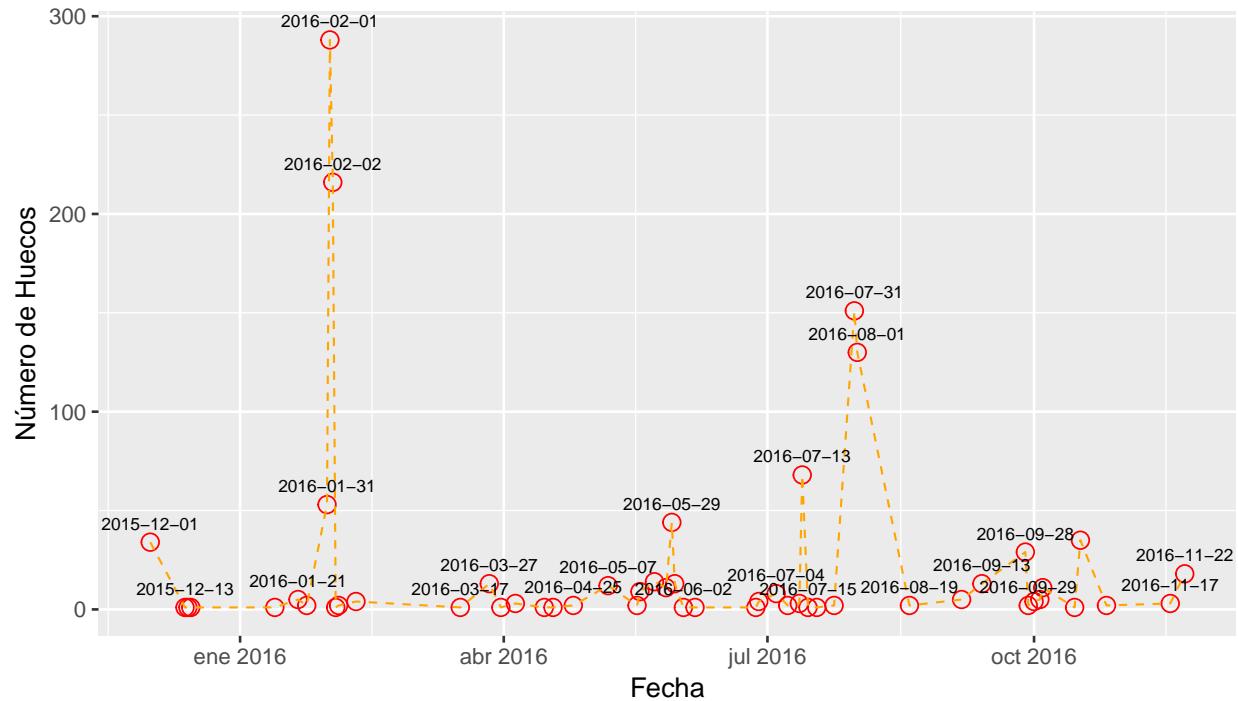


Figura 4: Datos faltantes. Huecos Globales por Fecha

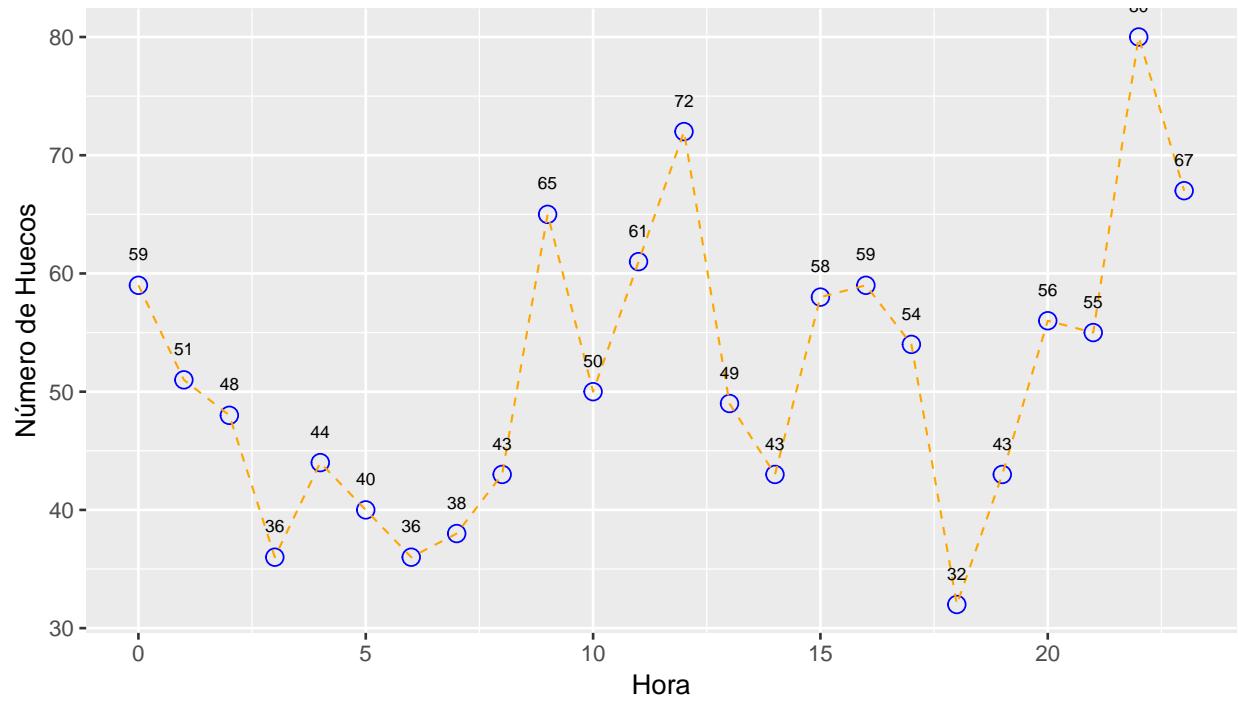


Figura 5: Datos faltantes. Huecos Globales por Hora

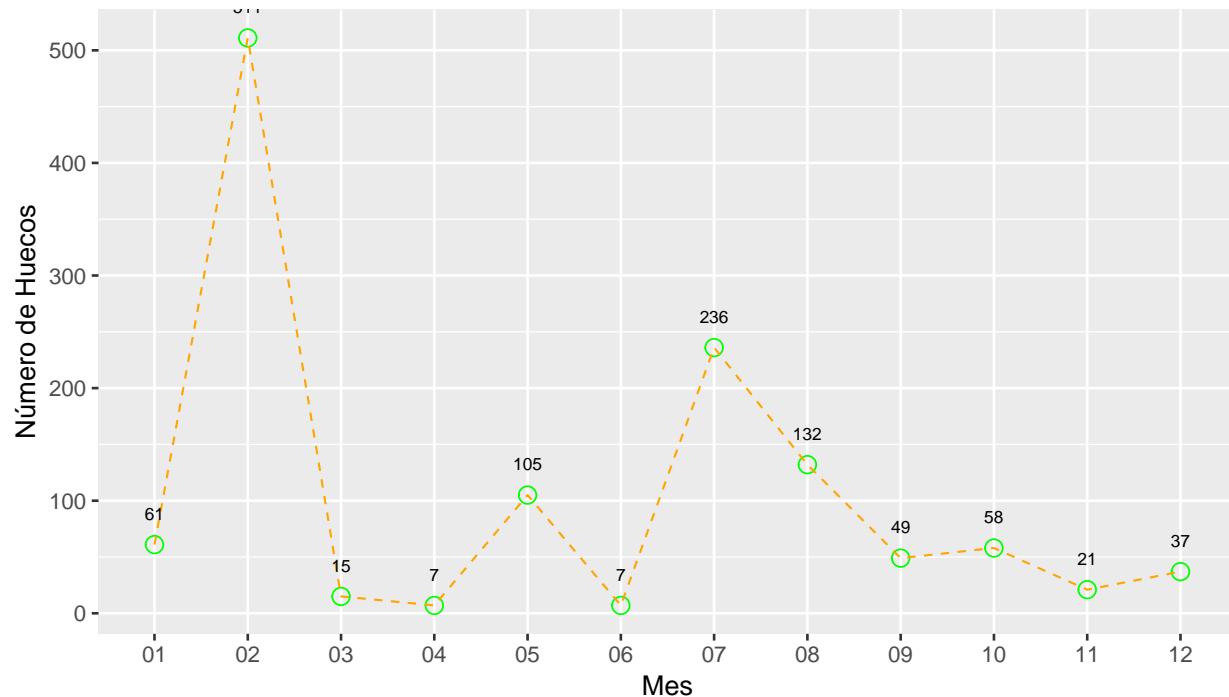


Figura 6: Datos faltantes. Huecos Globales por Mes

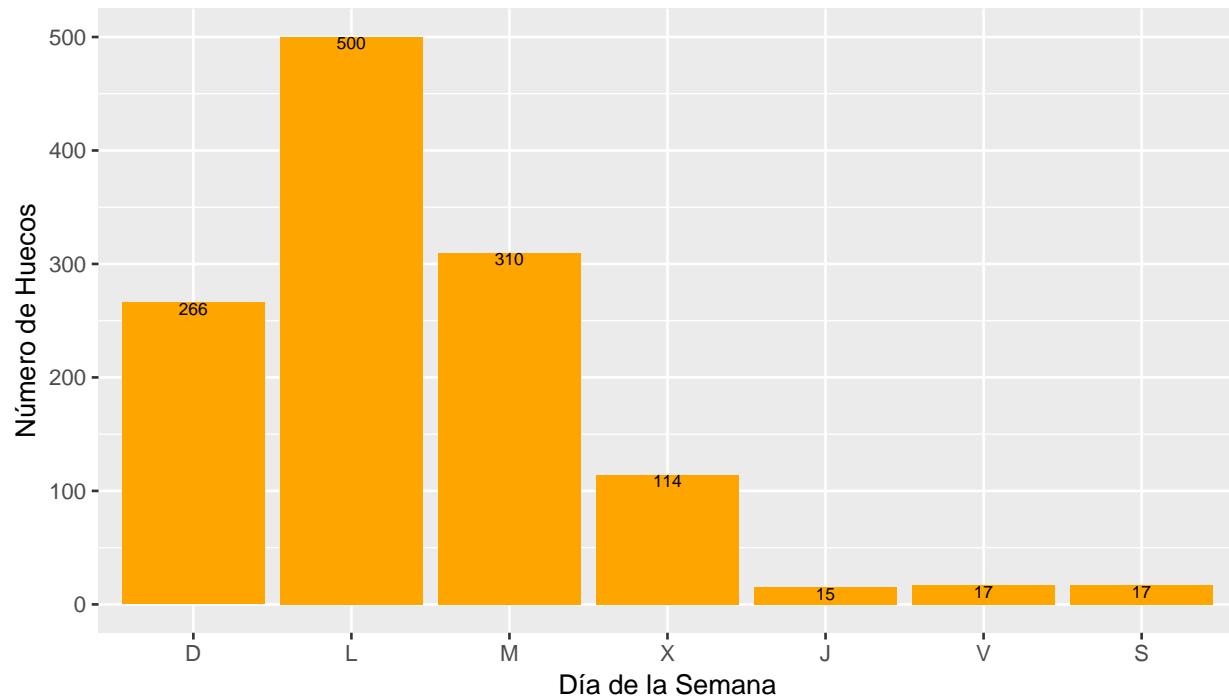


Figura 7: Datos faltantes. Huecos Globales por Día de la Semana

4. Datos anómalos

Para el análisis de datos anómalos se utiliza la estructura original de datos en la BD *seidata* en la que también se incluyen algunas vistas materializadas (tablas tmp) para facilitar un acceso más rápido a algunos datos.

Entre los datos anómalos se consideran las siguientes situaciones:

- a) Número de estacionamientos disponibles mayor que operativos
- b) Número de bicicletas disponibles mayor que estacionamientos operativos
- c) Suma de estacionamientos disponibles y bicicletas disponibles mayor que el número de estacionamientos operativos.
- d) Suma de estacionamientos disponibles y bicicletas disponibles menor que el número de estacionamientos operativos.

Se codifican dichas situaciones en la tabla *seidata* en el campo ok con los siguientes valores: a) -> ok = 3 b) -> ok = 4 c) -> ok = 5 d) -> ok = 6

Cuadro 3: Resumen de datos anómalos

ok_1	ok_2	ok_3	ok_4	ok_5	ok_6	TotOK_2_6	Total
22094898	2538	954	965	5728	4367735	4377920	26472818

Valor ok	Descripción
ok_1	Sin incidencia aparente
ok_2	Dato duplicado
ok_3	Estacionamientos disponibles > Est. operativos
ok_4	Bicicletas disponibles > Est. operativos
ok_5	Estacionamientos + Bicicletas disponibles > Est. operativos
ok_6	Estacionamientos + Bicicletas disponibles < Est. operativos

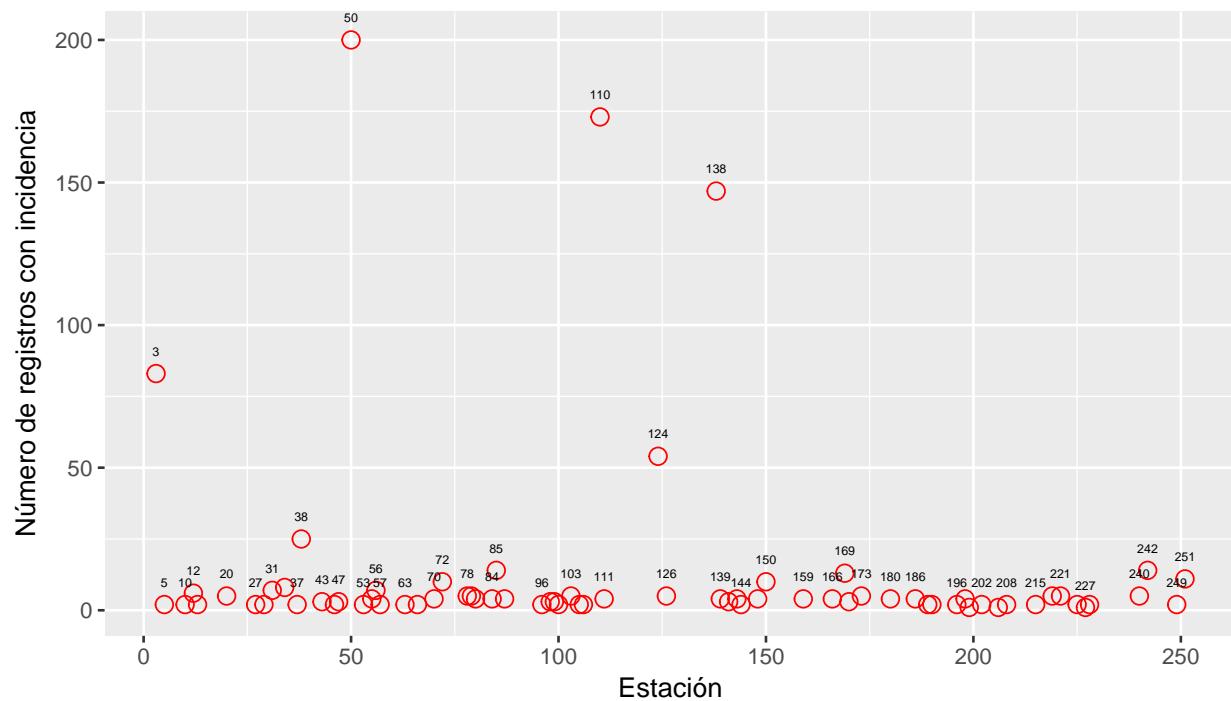


Figura 8: Datos anómalos. Estacionamientos disponibles > Est. Operativos

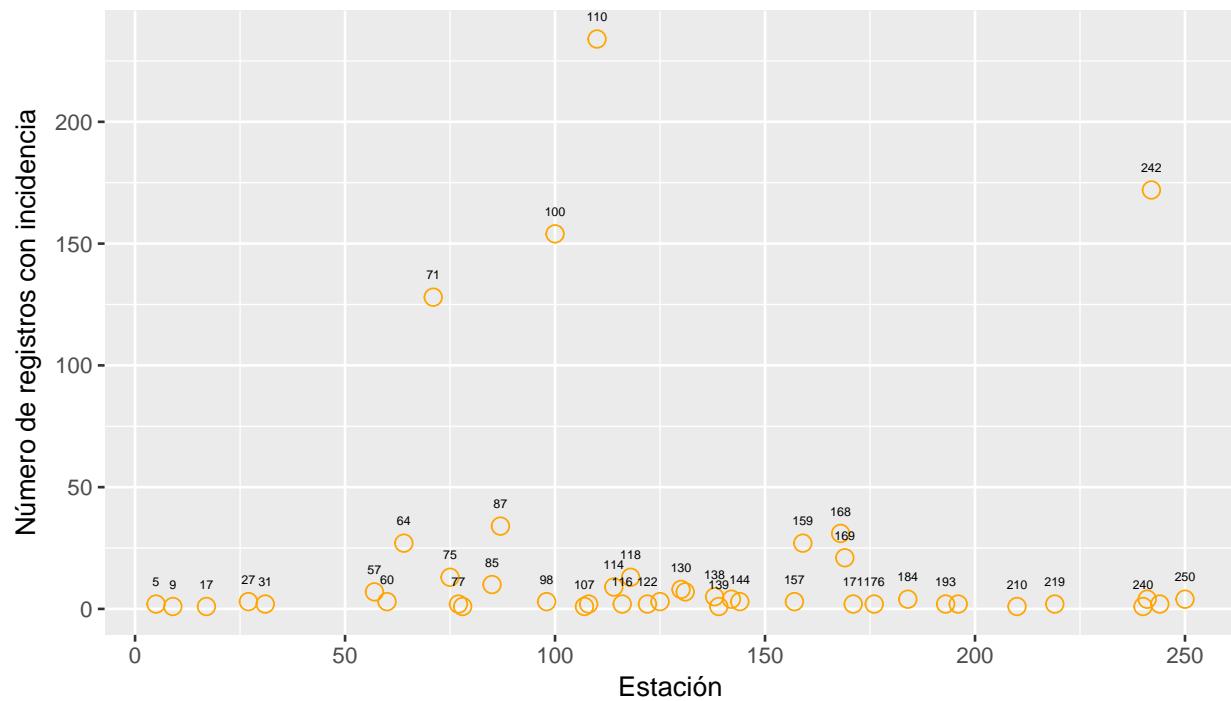


Figura 9: Datos anómalos. Bicicletas disponibles > Est. Operativos

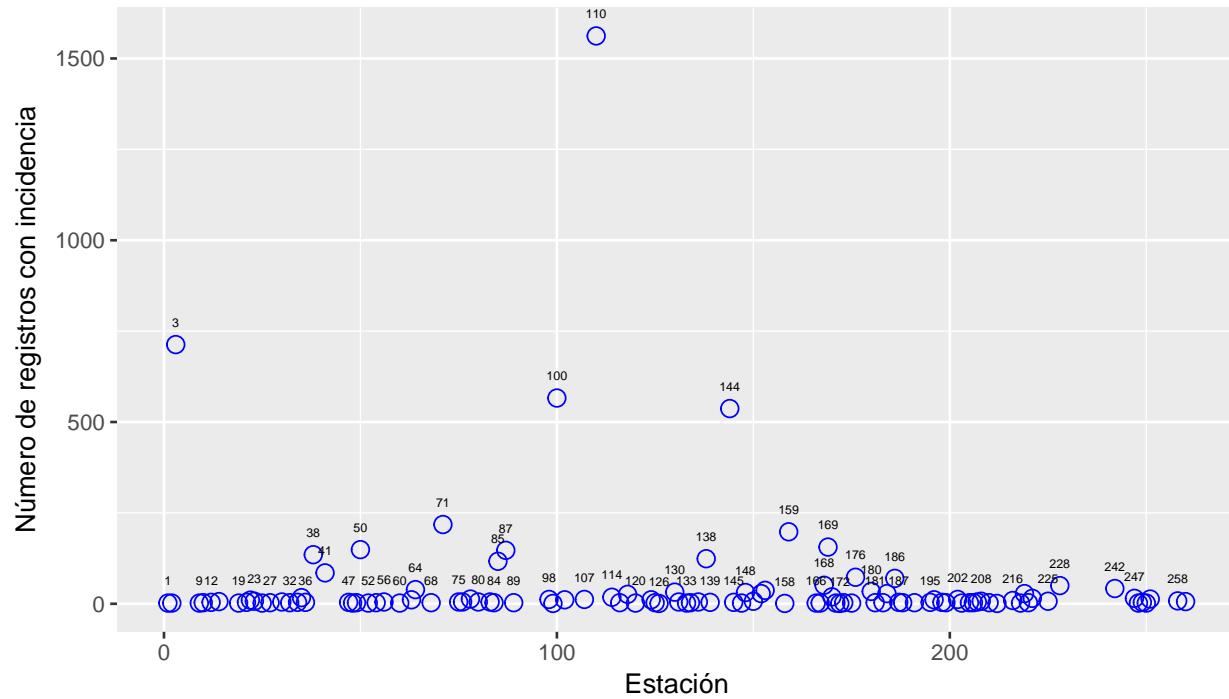


Figura 10: Datos anómalos. Estacionamientos + Bicicletas disponibles > Est. Operativos

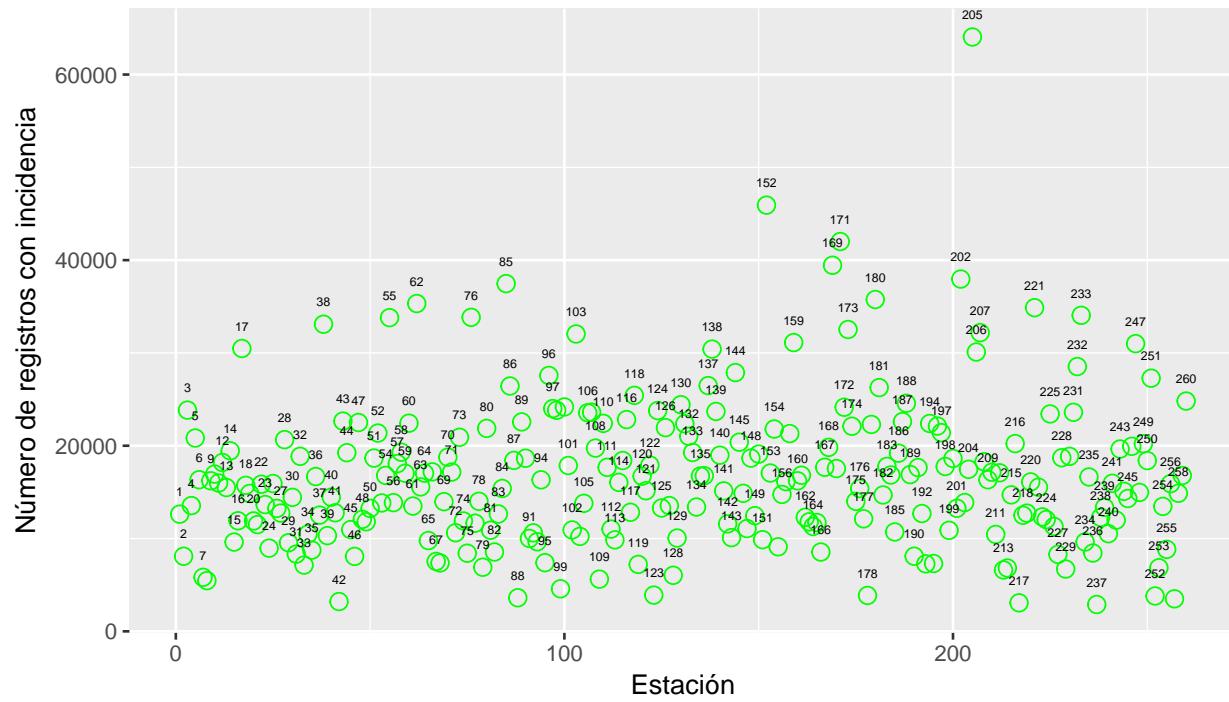


Figura 11: Datos anómalos. Estacionamientos + Bicicletas disponibles < Est. Operativos

5. Análisis de datos válidos globales

```
##          p5min nn  ss  sb
## 1 2015-12-01 00:00:01 211 2212 1797
## 2 2015-12-01 00:05:01 211 2210 1799
## 3 2015-12-01 00:10:01 211 2209 1800
## 4 2015-12-01 00:15:01 211 2198 1811
## 5 2015-12-01 00:20:01 211 2198 1811
## 6 2015-12-01 00:25:01 211 2200 1809
```

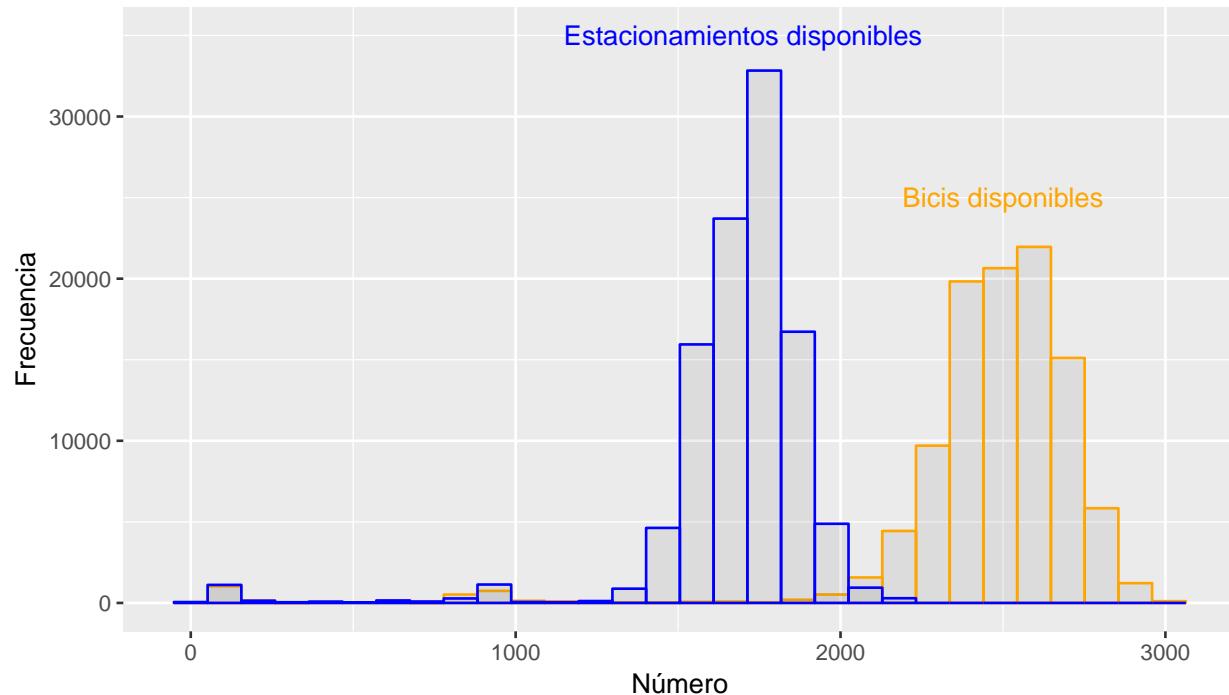


Figura 12: Datos válidos globales. Distribución de Estacionamientos y Bicis disponibles.

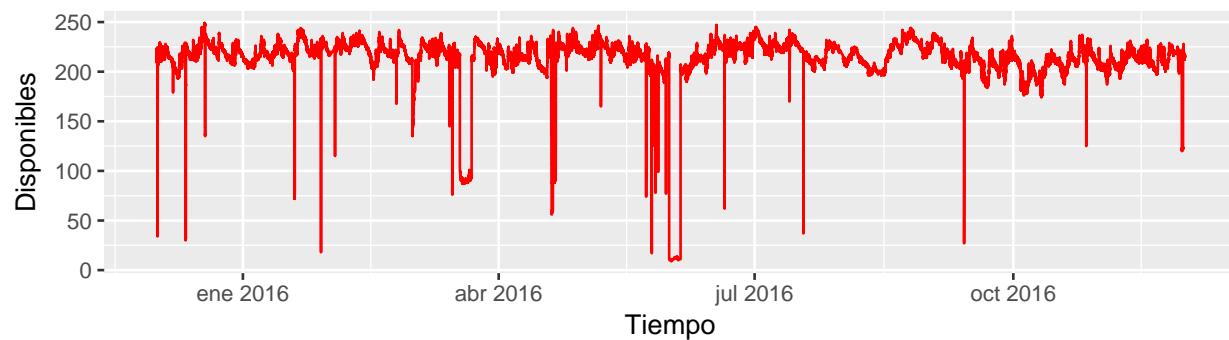


Figura 13: Datos válidos globales. Número de estaciones disponibles.

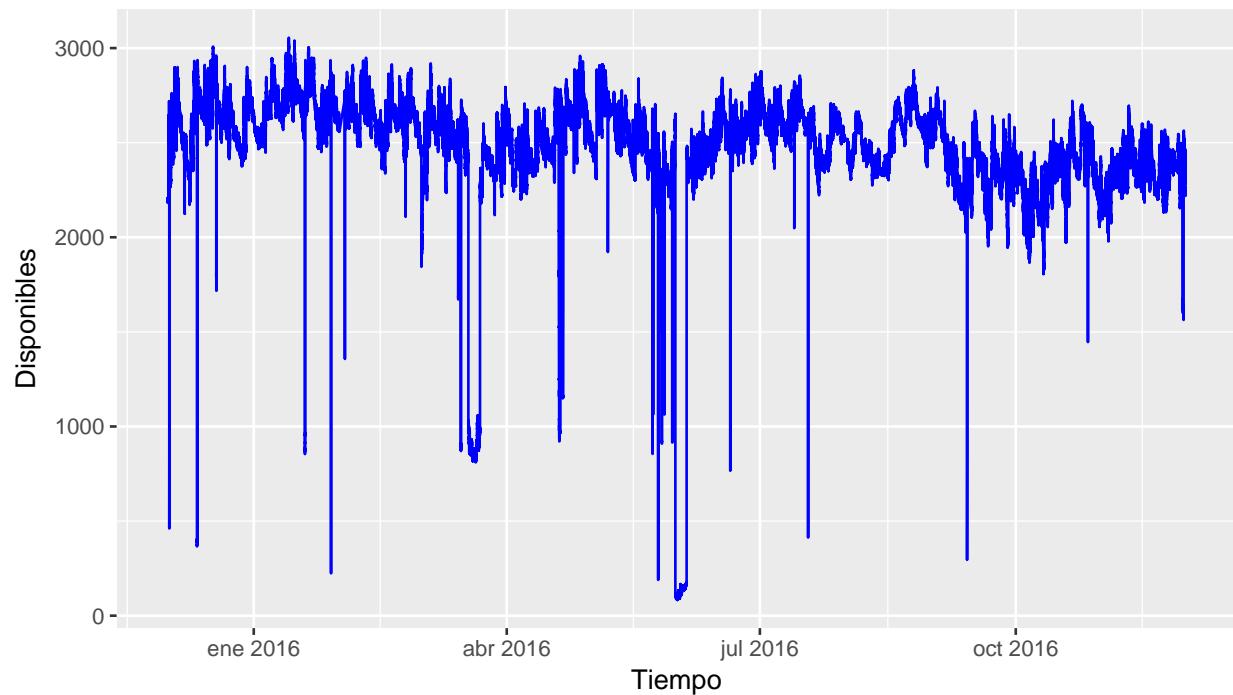


Figura 14: Datos válidos globales. Estacionamientos disponibles.

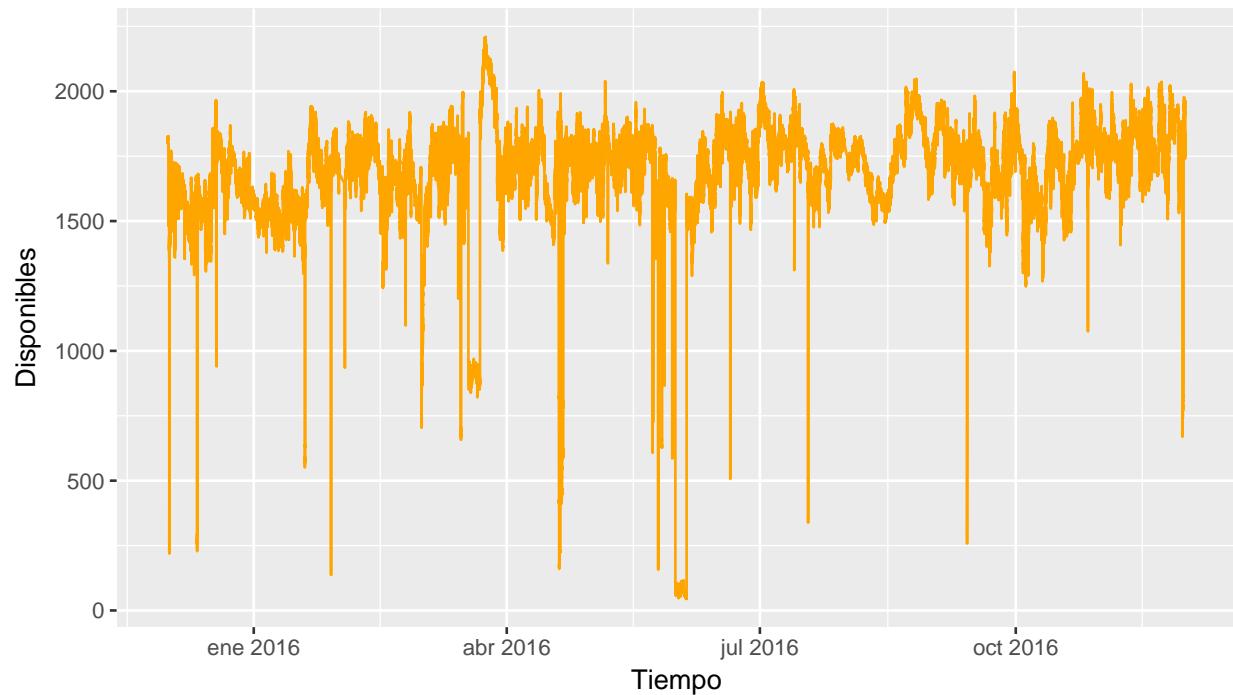


Figura 15: Datos válidos globales. Bicis disponibles.

El número de estacionamientos disponibles es asimilable al número de bicis circulantes, y a la vista del gráfico, parece existir un número de bicis circulantes muy superior al de bicis disponibles a lo largo de todo el periodo. Fenómeno con una aparente tendencia a su reducción.

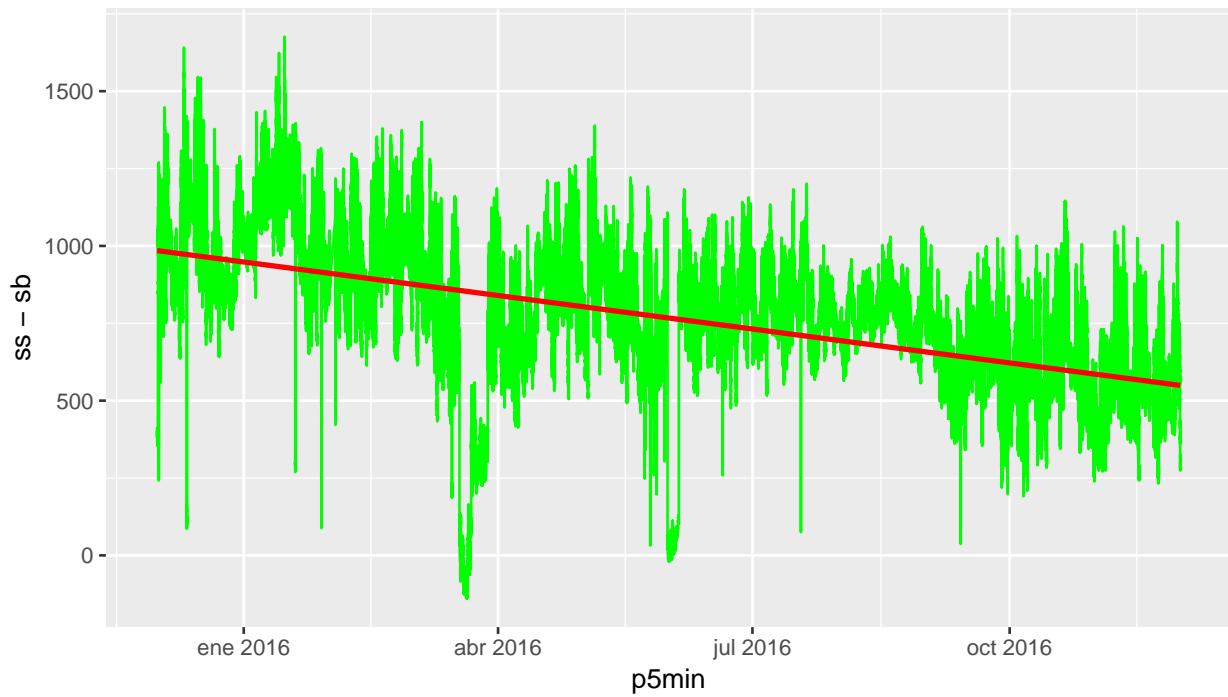


Figura 16: Datos válidos globales. Diferencia Bicis circulantes y Bicis disponibles.

```
## $x
## [1] "Tiempo"
##
## $y
## [1] "Diferencia"
##
## attr(,"class")
## [1] "labels"
```

5.1. Análisis según días de la semana

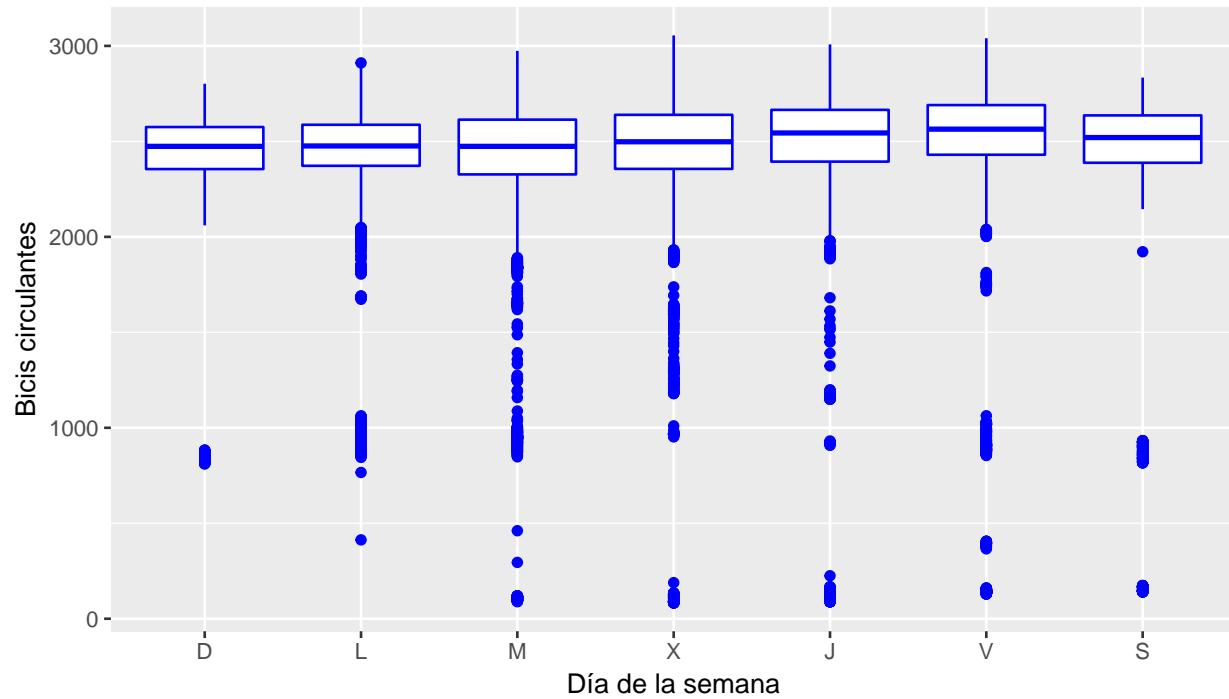


Figura 17: Datos válidos globales. Bicis circulantes según día de la semana.

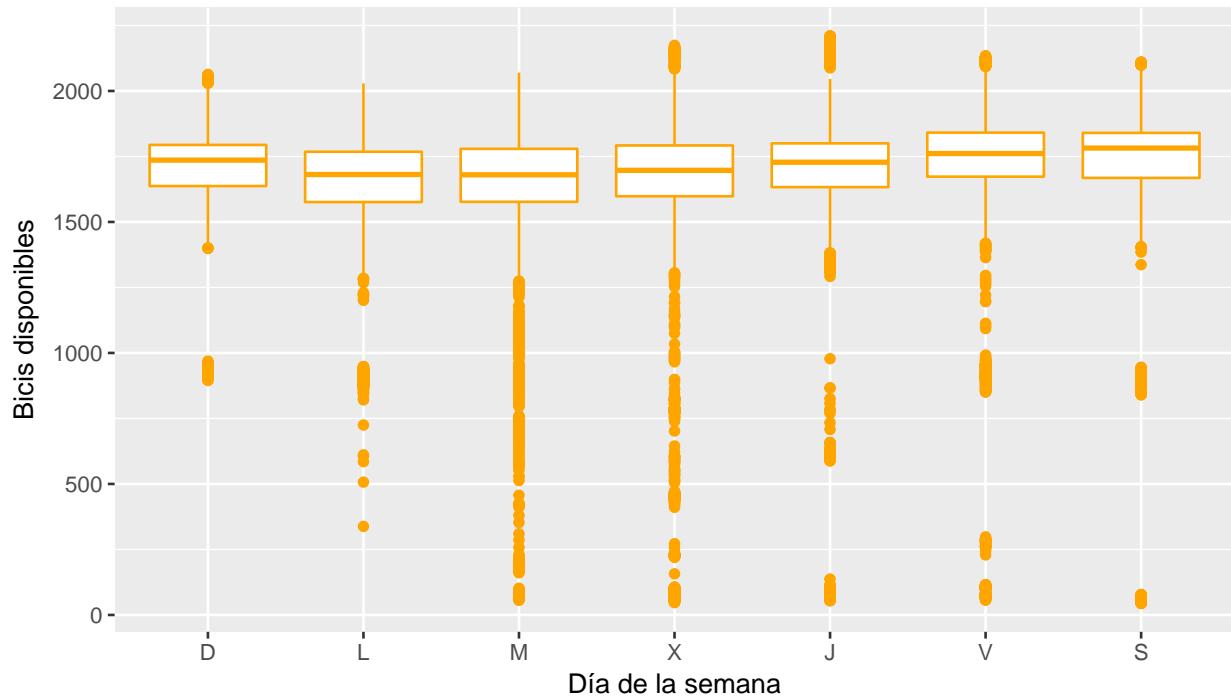


Figura 18: Datos válidos globales. Bicis disponibles según día de la semana.

Cuadro 5: Bicis circulantes por día de la semana. Estadística básica.

dsem	median	mean	min	max	sd
M	2474	2428.767	89	2974	333.3981
D	2474	2433.081	811	2802	263.3739
L	2476	2448.293	413	2911	269.6886
X	2498	2426.666	83	3055	416.5099
S	2520	2455.694	140	2834	373.5232
J	2544	2478.600	89	3008	396.7103
V	2564	2486.623	129	3040	438.3324

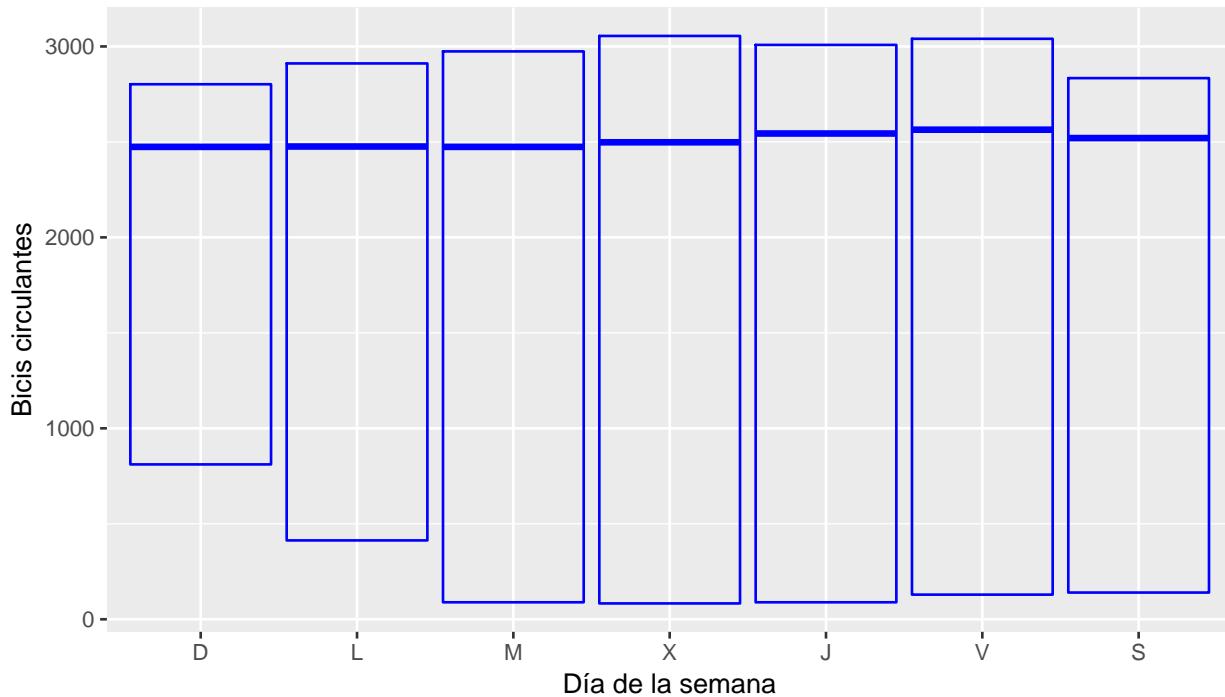


Figura 19: Datos válidos globales. Bicis circulantes por día de la semana. Estadística básica. Mediana, Máximo y Mínimo.

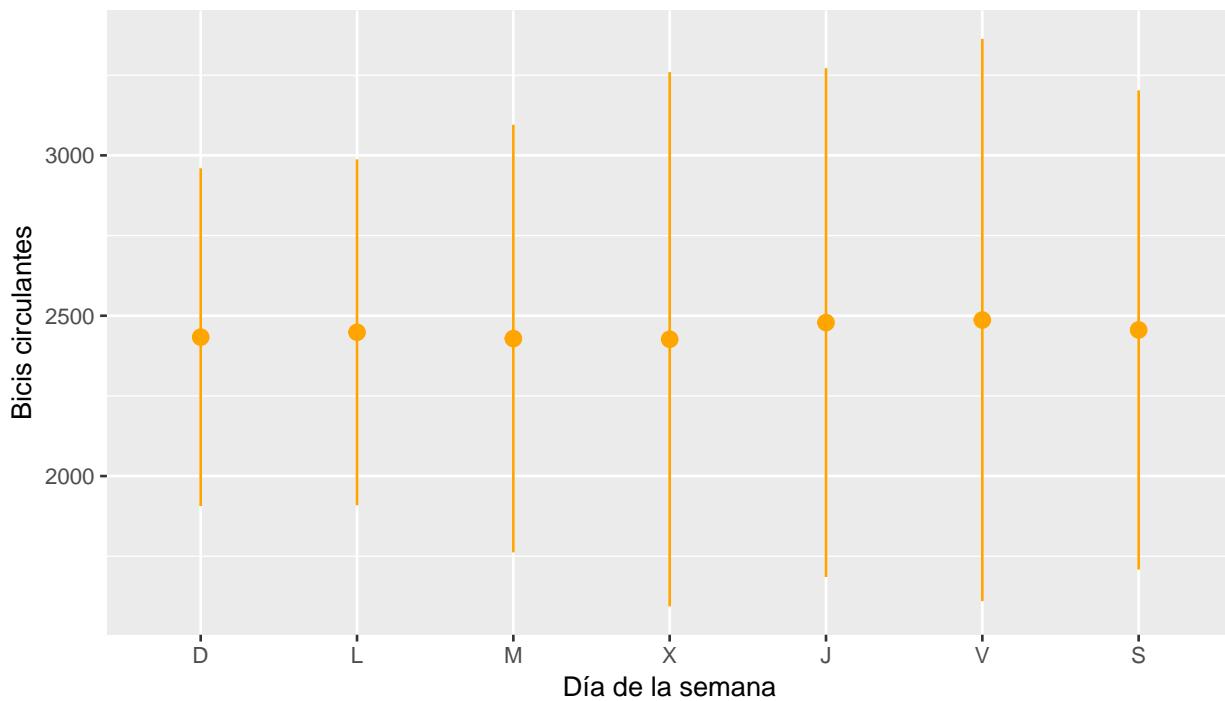


Figura 20: Datos válidos globales. Bicis circulantes por día de la semana. Media +/- 2 · Desviación

5.2. Análisis según hora del día

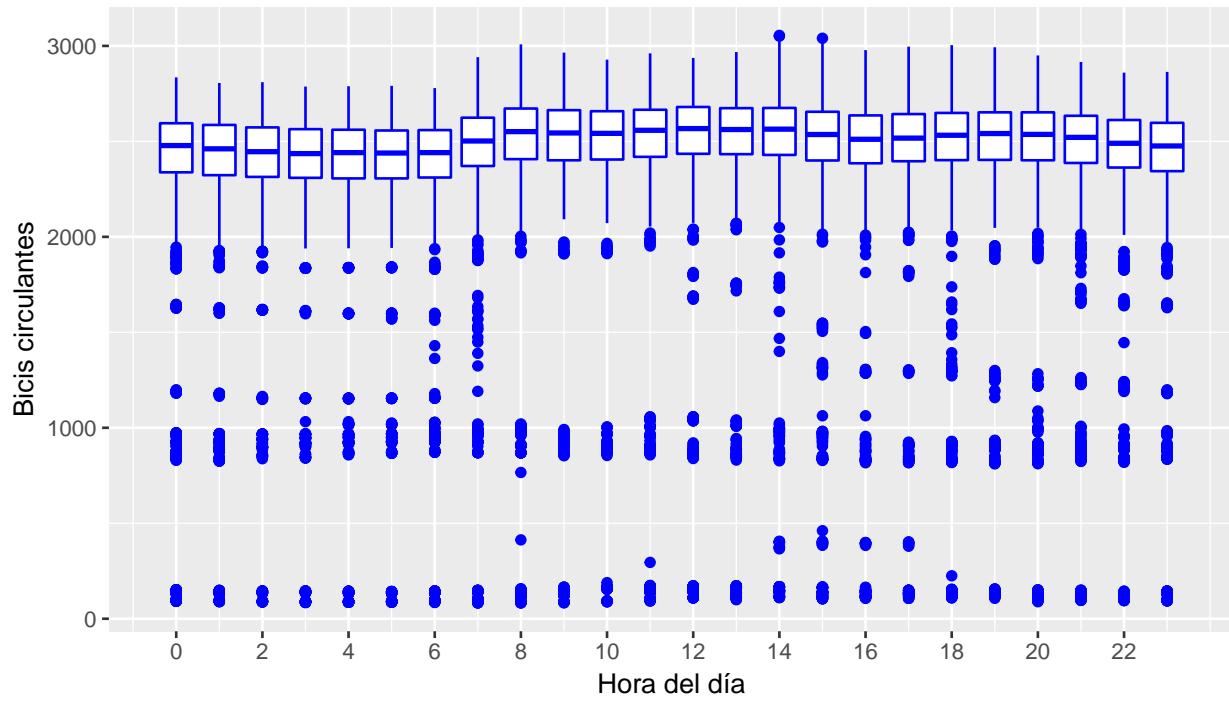


Figura 21: Datos válidos globales. Bicis circulantes según hora del día.

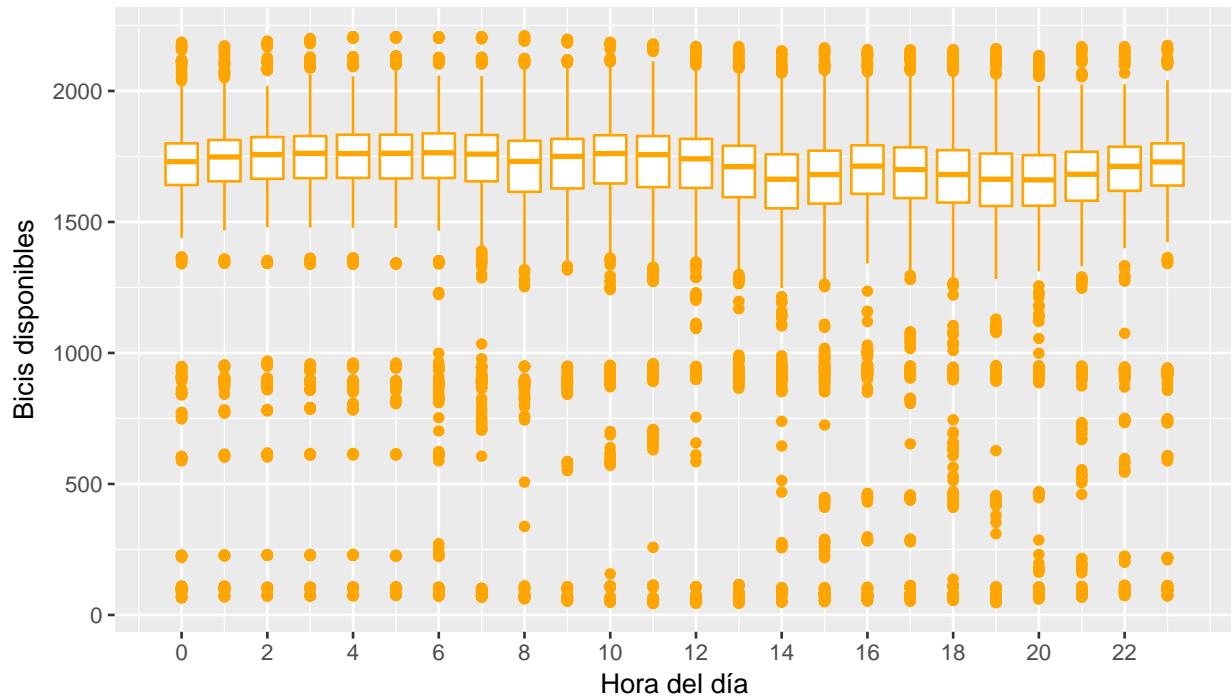


Figura 22: Datos válidos globales. Bicis disponibles según hora del día.

Cuadro 6: Bicis circulantes por hora del día. Estadística básica.

hora	median	mean	min	max	sd
3	2436.0	2384.347	85	2787	351.2637
5	2439.0	2377.993	87	2791	355.3222
4	2441.0	2377.850	86	2789	356.7811
6	2441.0	2380.828	85	2779	354.0692
2	2446.0	2389.804	88	2810	352.6853
1	2461.0	2399.390	89	2806	355.4711
23	2476.0	2414.556	93	2864	358.3391
0	2478.0	2412.367	91	2835	356.0342
22	2490.0	2430.536	95	2860	361.4458
7	2502.0	2442.667	83	2941	354.2777
16	2511.0	2456.665	107	2978	372.3386
17	2517.0	2466.635	107	2996	365.0067
21	2521.0	2458.150	96	2916	362.7715
18	2532.0	2472.911	109	3004	366.3834
15	2536.0	2474.662	104	3040	395.2954
20	2536.5	2479.051	89	2950	362.1650
19	2541.0	2482.074	107	2993	359.1806
10	2542.0	2487.485	89	2928	356.4582
9	2544.0	2492.910	84	2965	353.8283
8	2551.0	2503.843	83	3008	355.1563
11	2558.0	2500.450	92	2961	360.5419
13	2562.0	2515.749	100	2968	347.3083
14	2564.0	2511.131	110	3055	373.7291
12	2567.0	2515.499	109	2937	351.8415

hora	median	mean	min	max	sd
------	--------	------	-----	-----	----

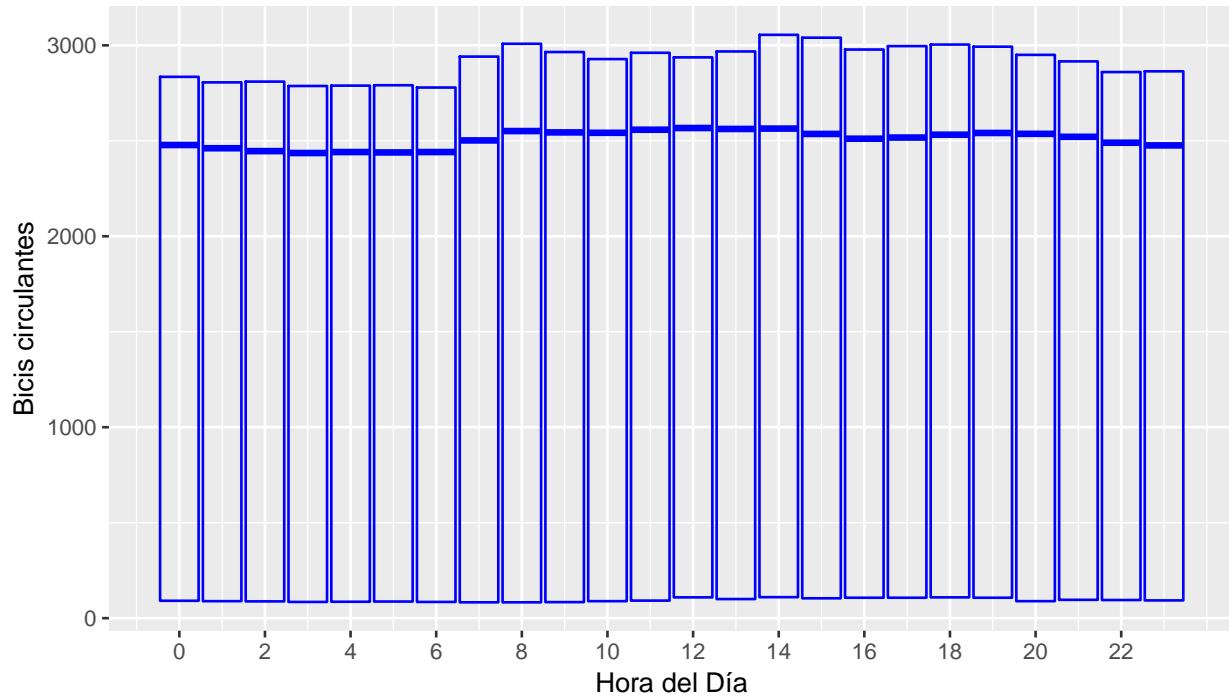


Figura 23: Datos válidos globales. Bicis circulantes por hora del día. Estadística básica. Mediana, Máximo y Mínimo

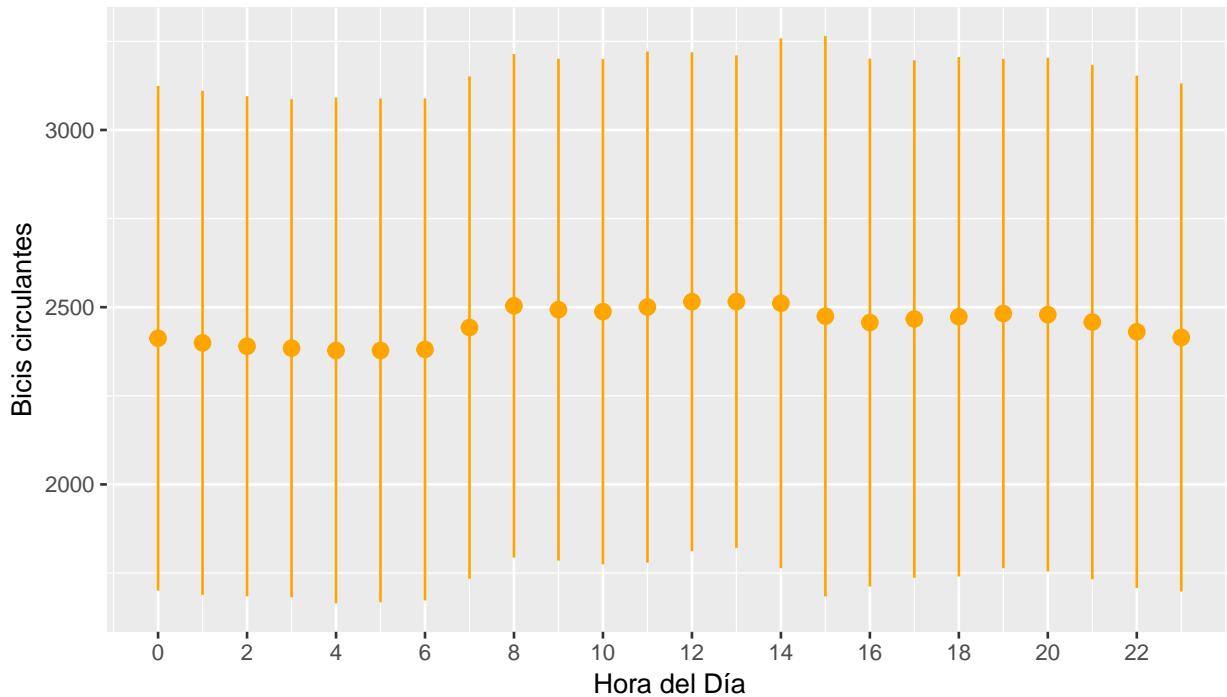


Figura 24: Datos válidos globales. Bicis circulantes por hora del día. Estadística básica. Media $+$ $- 2 \cdot$ Desviación

5.3. Análisis según hora del día y día de la semana

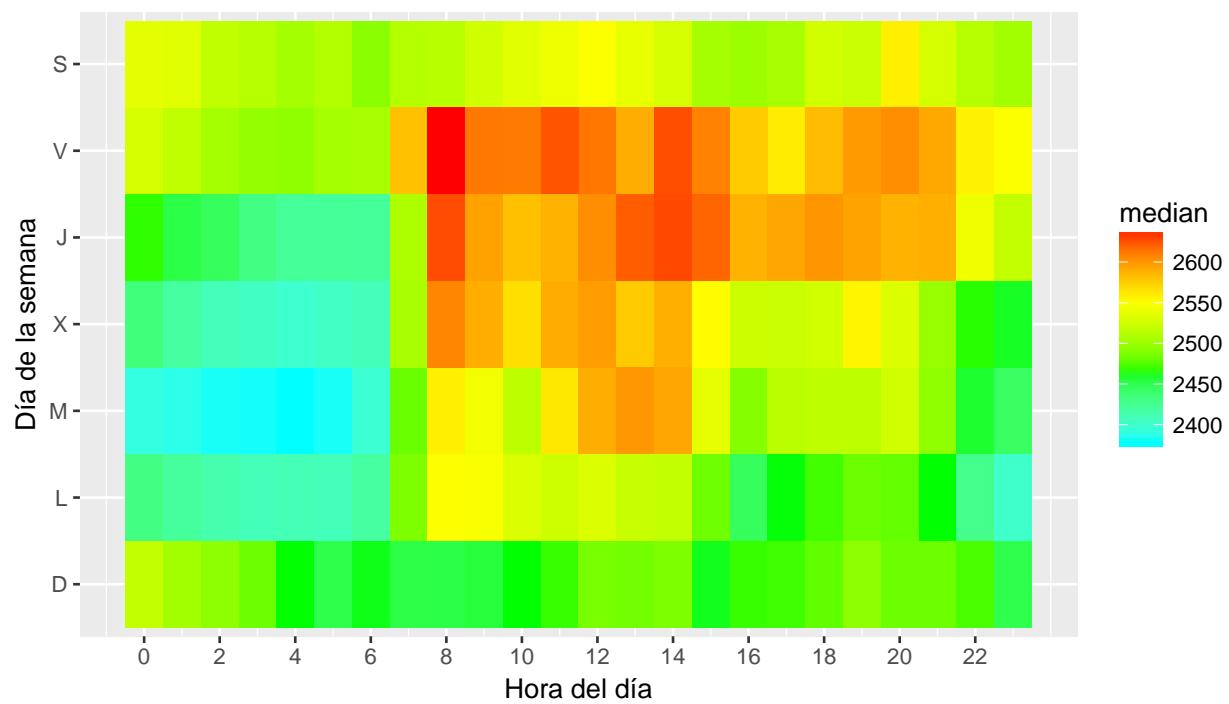


Figura 25: Datos válidos globales. Bicis circulantes según hora del día y día de la semana.

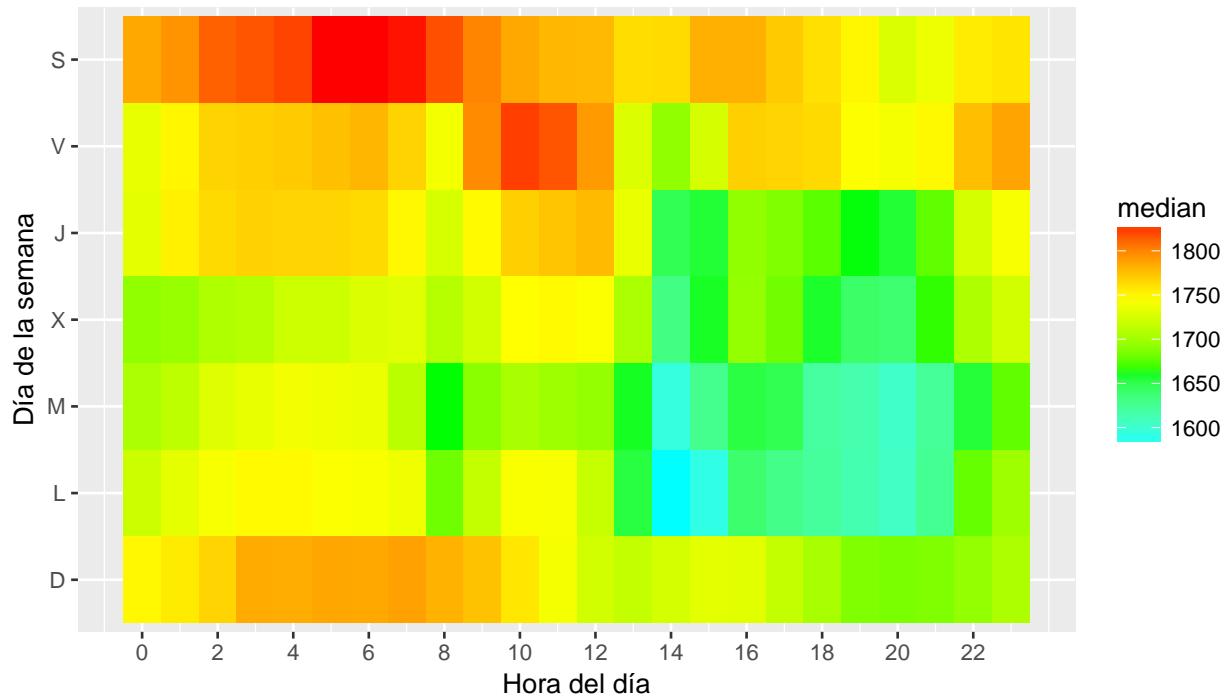


Figura 26: Datos válidos globales. Bicis disponibles según hora del día y día de la semana.

6. Análisis de datos válidos por estaciones

Presentamos a continuación los resultados visuales del análisis de los datos válidos de las estaciones.

6.1. Análisis de correlación entre estaciones

Representamos la matriz de correlación (Pearson) entre estaciones para la variable número de estacionamientos disponibles por estación.

Segregamos la matriz de correlación en un dataframe con todos los pares y el valor de correlación para su tratamiento posterior como grafo (con nodos geoposicionados).

```
## # A tibble: 67,600 x 5
##   from     to    value  Var1  Var2
##   <int> <int>    <dbl> <fctr> <fctr>
## 1     1      1  1.000000000  s1     s1
## 2     2      1  0.08251733  s2     s1
## 3     3      1  0.04781273  s3     s1
## 4     4      1 -0.18770758  s4     s1
## 5     5      1 -0.20230016  s5     s1
## 6     6      1 -0.17154154  s6     s1
## 7     7      1 -0.31601164  s7     s1
## 8     8      1 -0.19564438  s8     s1
## 9     9      1 -0.21507347  s9     s1
## 10   10     1  0.20802432  s10    s1
## # ... with 67,590 more rows
```

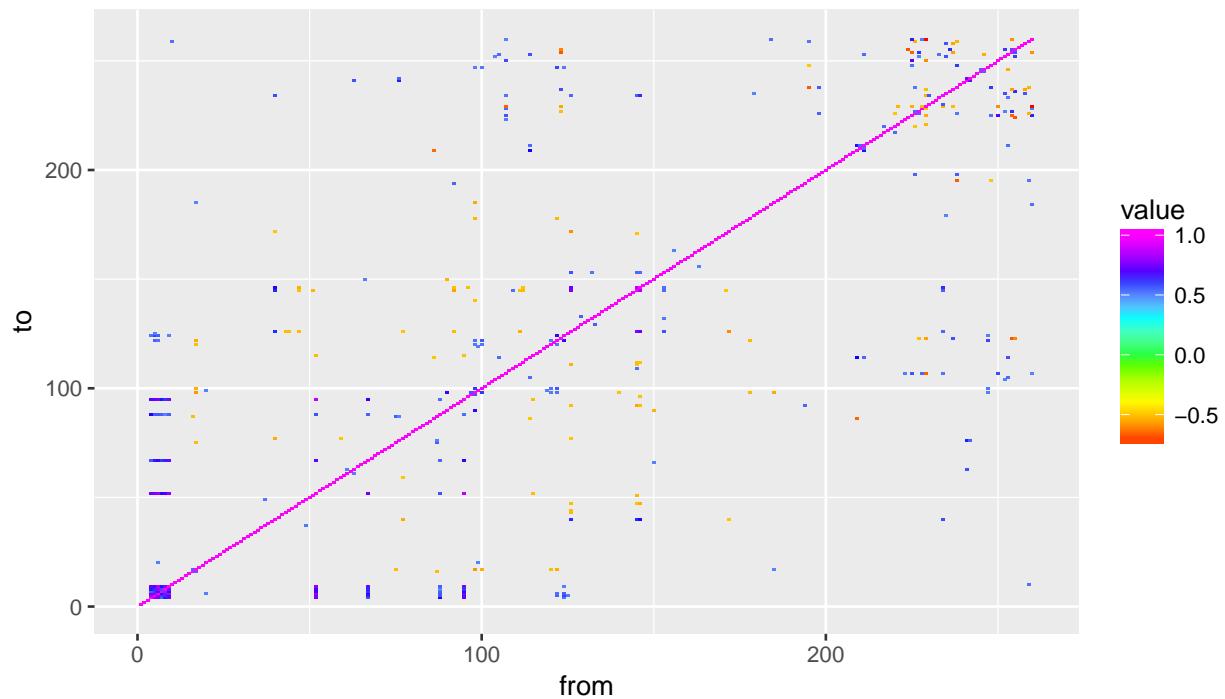


Figura 27: Datos válidos estaciones. Matriz de correlación ($|\text{corr}|>0.5$) entre estaciones.

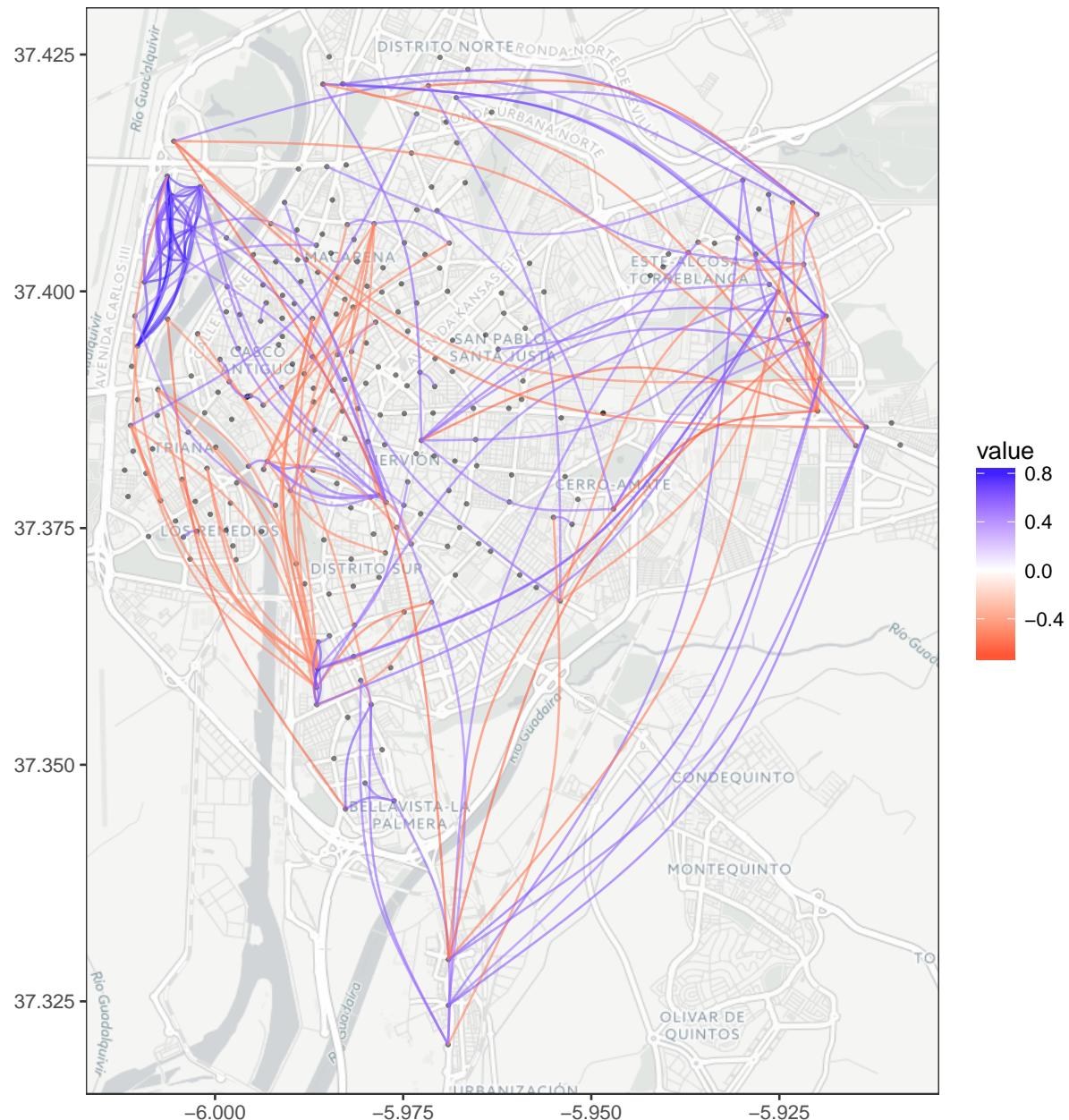


Figura 28: Datos válidos estaciones. Grafo espacial de correlaciones $|\text{corr}| > 0.5$.

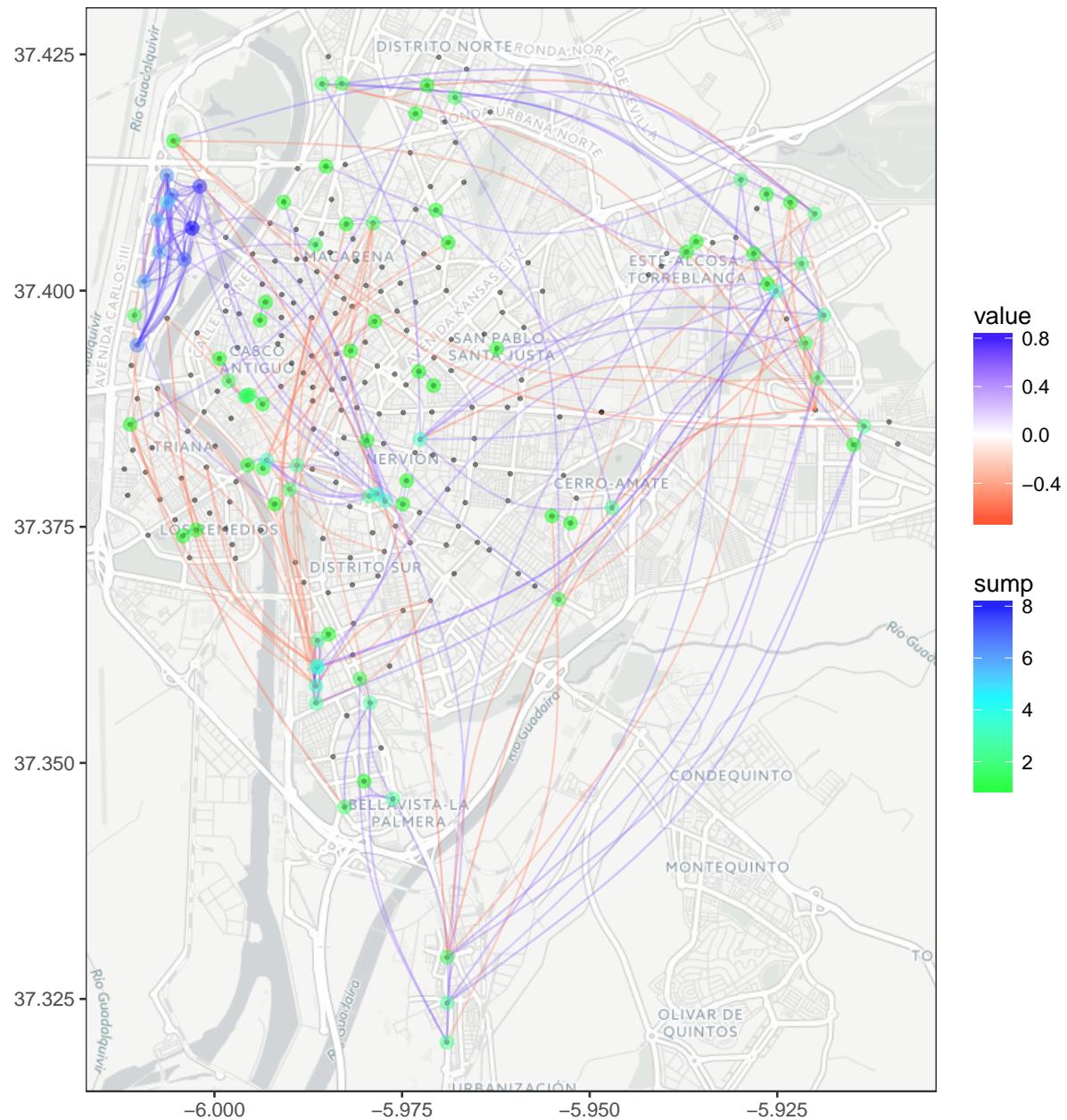


Figura 29: Datos válidos estaciones. Suma de correlaciones positivas por estación ($\text{corr} > 0.5$).

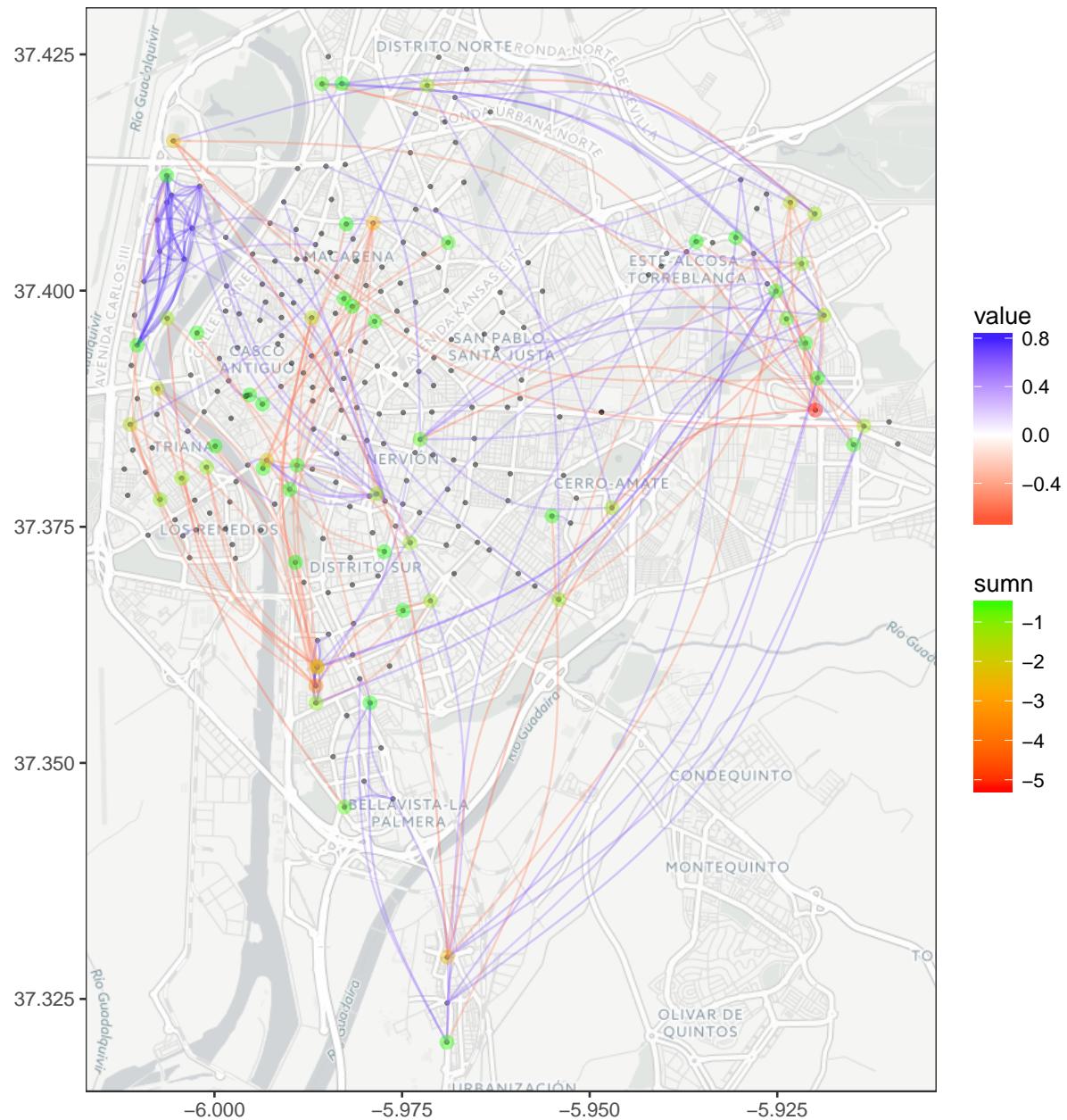


Figura 30: Datos válidos estaciones. Suma de correlaciones negativas por estación ($\text{corr} < -0.5$).

6.2. Clasificación de las estaciones

Utilizamos la matriz de correlación como base para la clasificación de las estaciones. Para ello en primer lugar convertimos los coeficientes de correlación en disimilaridades y éstas son tratadas como distancias.

```
##          Length Class  Mode
## merge      518   -none- numeric
## height     259   -none- numeric
## order      260   -none- numeric
## labels     260   -none- character
## method      1    -none- character
## call        2    -none- call
## dist.method 0    -none- NULL
```

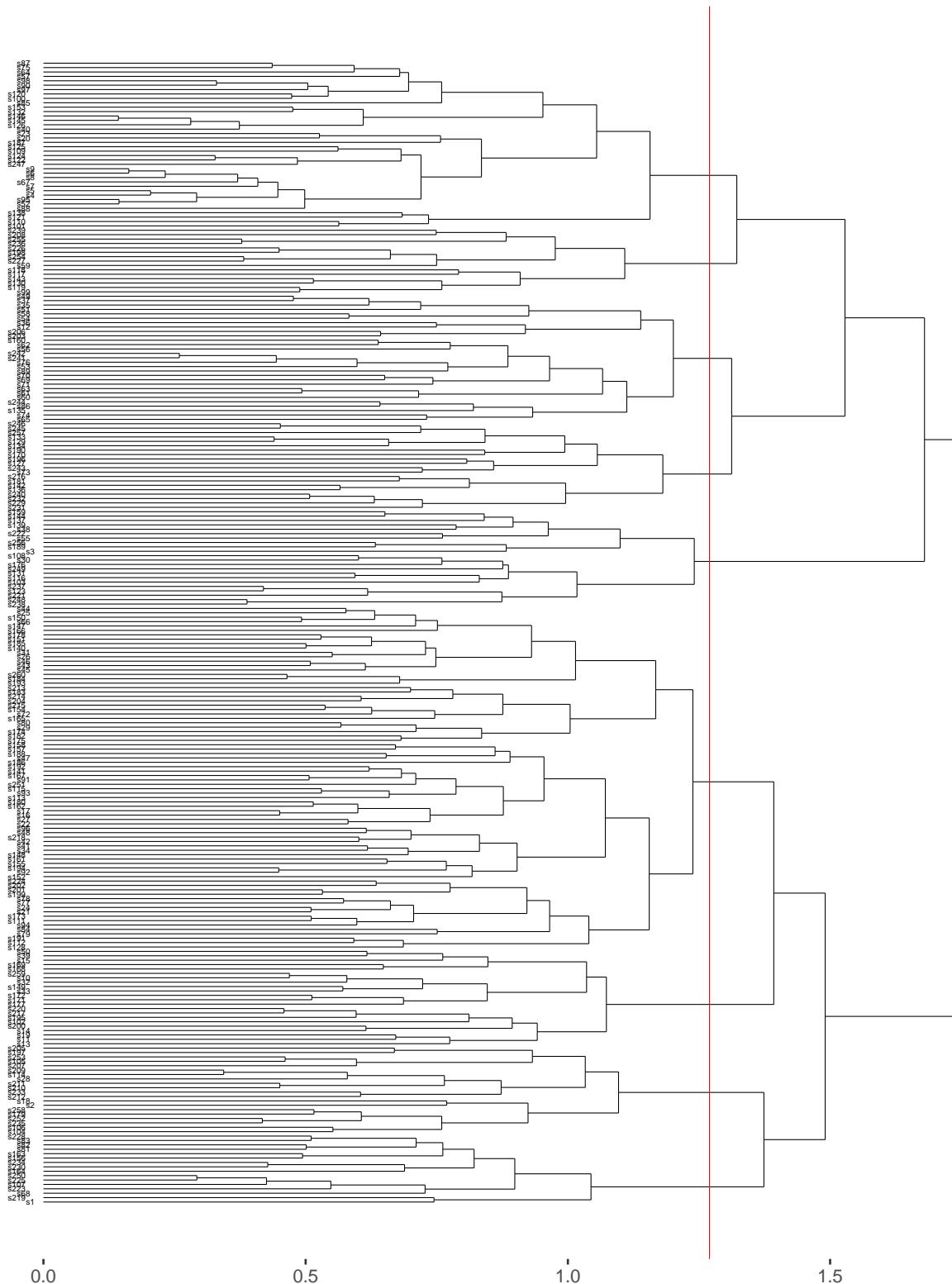


Figura 31: Datos válidos estaciones. Dendrograma de estaciones basado en correlación.

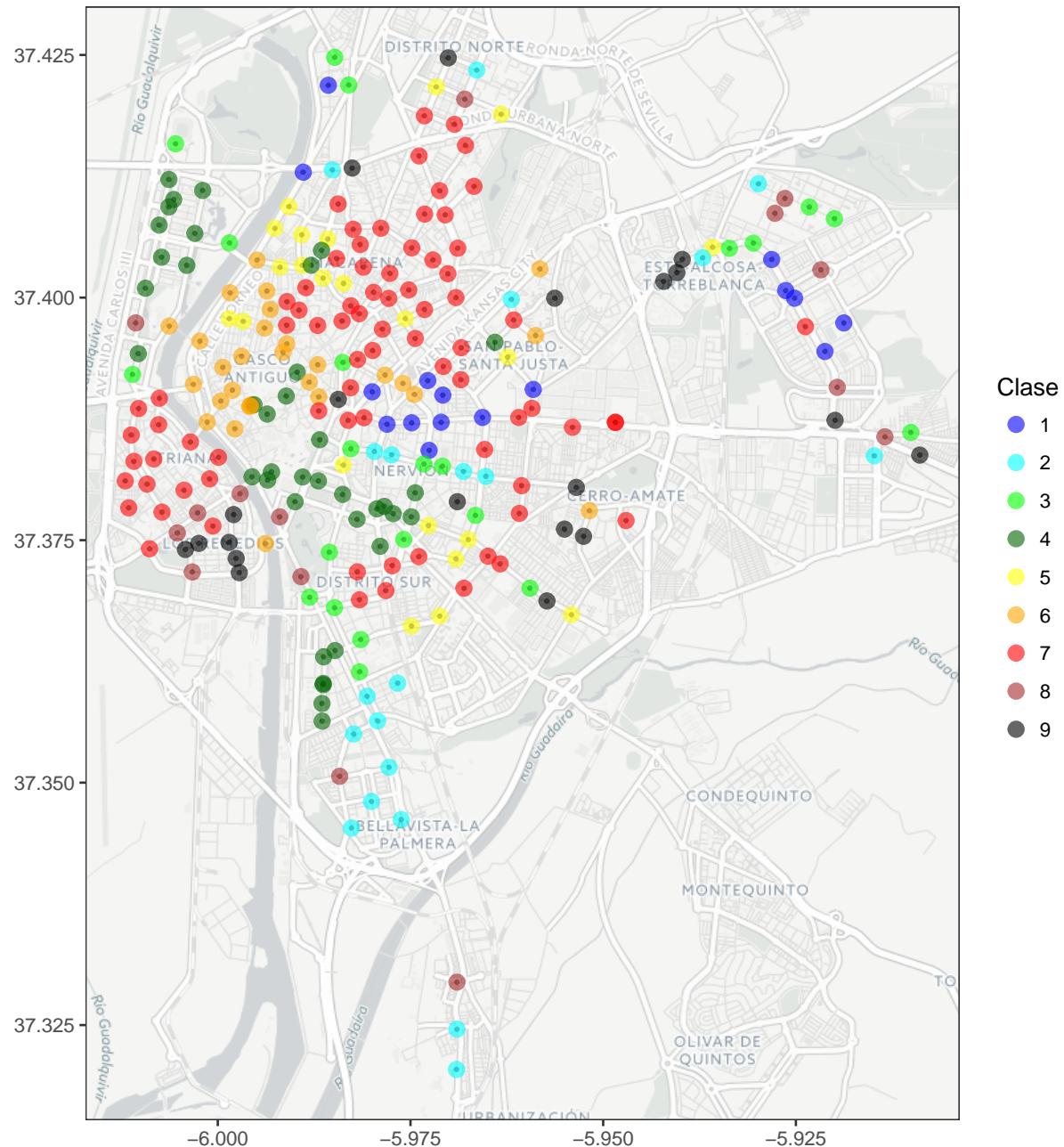


Figura 32: Datos válidos estaciones. Clasificación de estaciones.

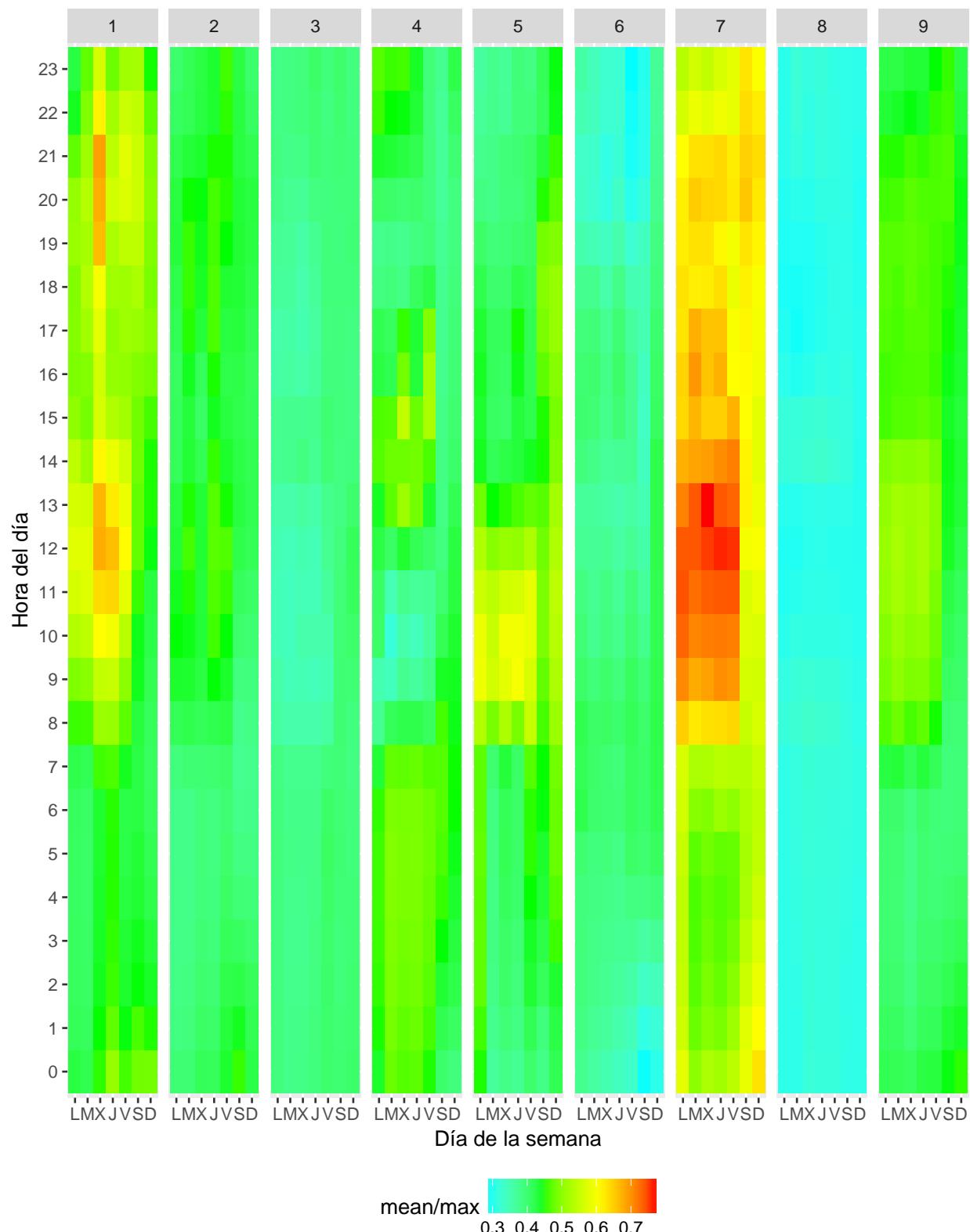
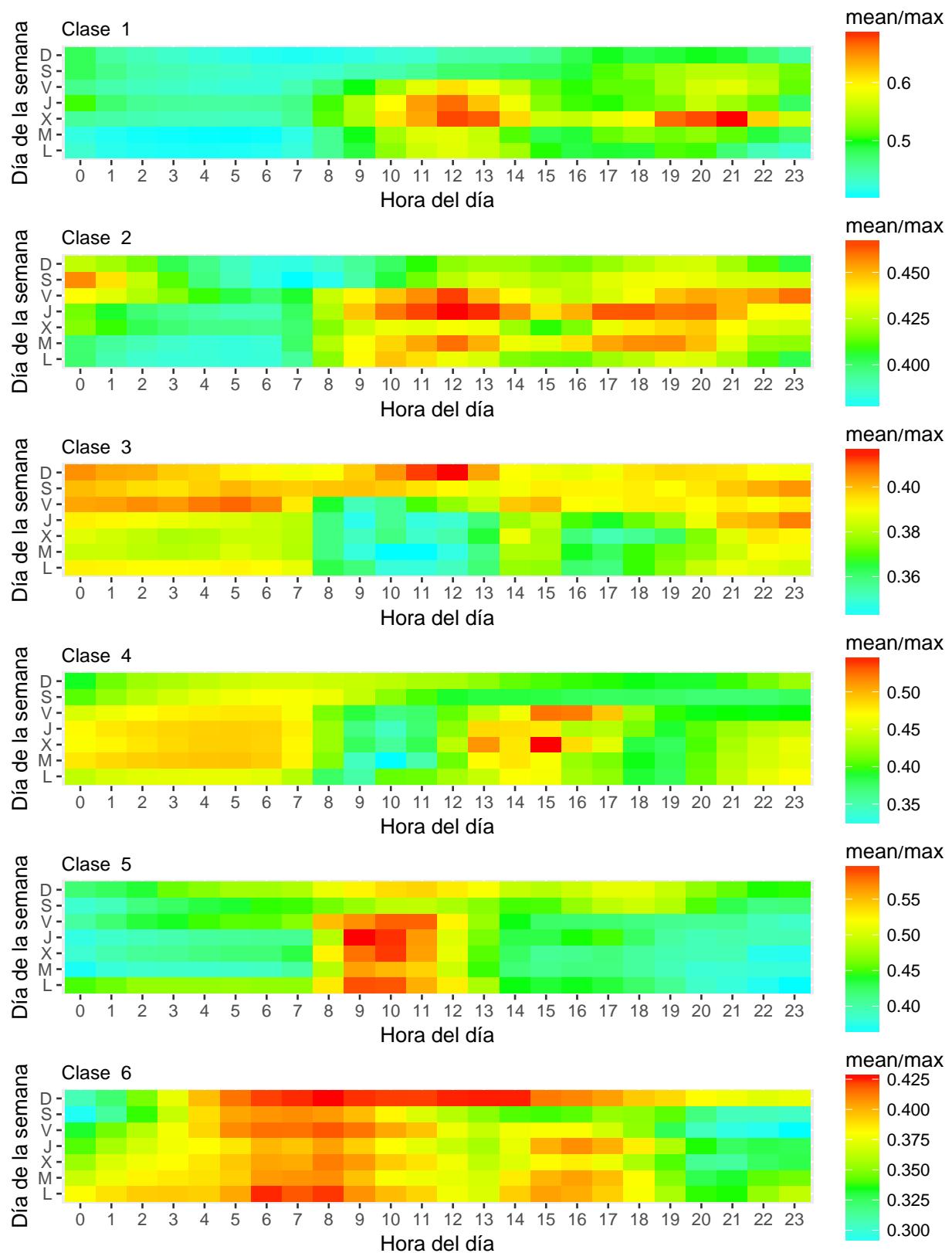


Figura 33: Datos estaciones. Estacionamientos disponibles por clase de estación, hora del día y día de la semana.



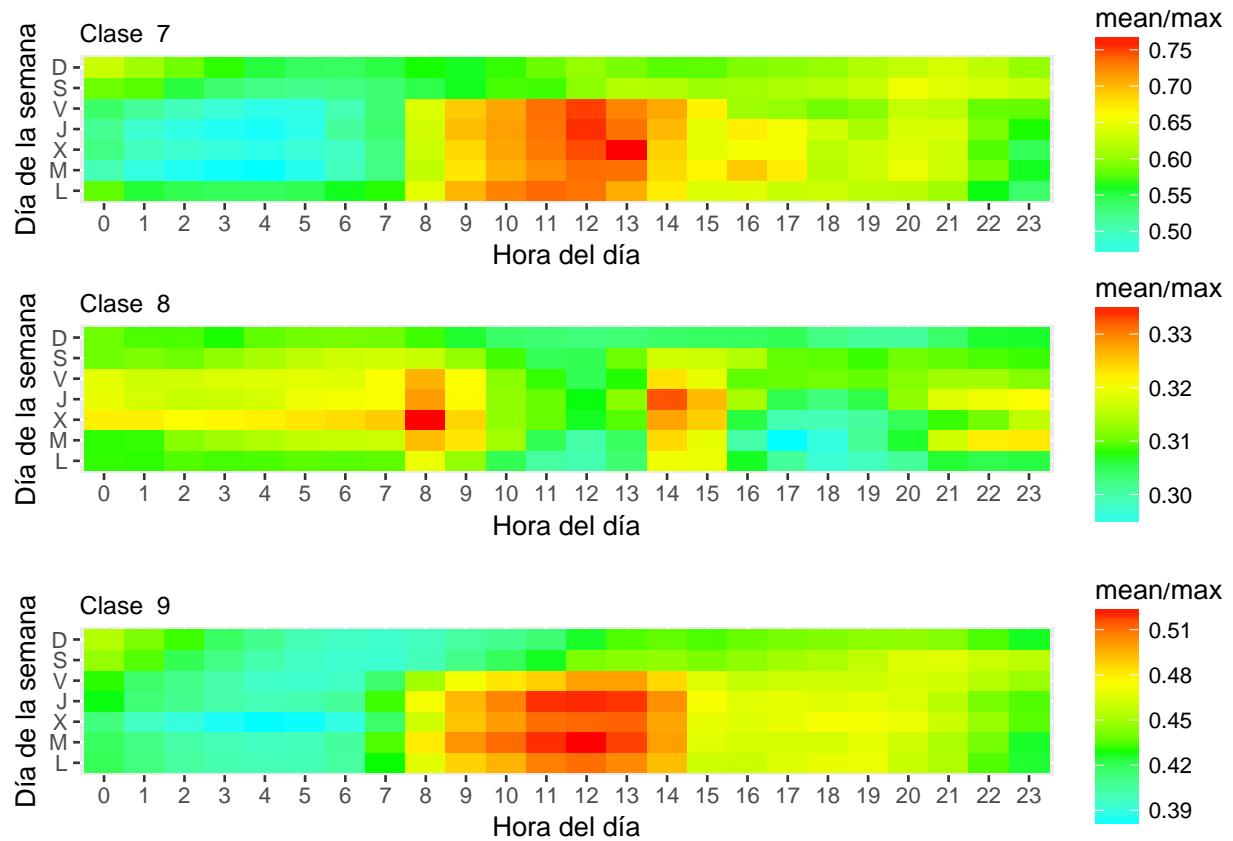


Figura 34: Datos estaciones. Estacionamientos disponibles por hora del día y día de la semana. Patrones por clase de estación.