

Proyecto final

Iván Arriola, Federico Miquelerena, Damián Rovetta

12-07-2023

Introducción

Este trabajo es más que descriptivo, la introducción es muy limitada

Esto es un análisis descriptivo de los datos del tráfico de Montevideo, Uruguay. Hemos tomado los registros desde enero de 2021 hasta mayo de 2023 y nuestro interés es saber el comportamiento de la velocidad y el volumen de tráfico (variables explicativas) dependiendo de varias variables que iremos desarrollando a lo largo de la investigación.

Datos

Descripción general de los datos

Todos los datos fueron sacados de Catalogo de Datos Abiertos de **gub.uy**. En particular, los datos elegidos son los siguientes:

- Conteo vehicular en las principales avenidas de Montevideo
- Velocidad promedio vehicular en las principales avenidas de Montevideo
- Ubicación de sensores de medición de conteo vehículos

Los tres dataset son mantenidos por la Intendencia de Montevideo.

Descripción de variables

Originalmente los datos vienen presentados de la siguiente forma:

Conteo vehicular en las principales avenidas de Montevideo

- **cod_detector**: Numérico - ID de la cámara que monitorea un carril específico para detectar vehículos.
- **id_carril**: Numérico - Número del carril monitoreado (1, 2, 3, ...).
- **fecha**: Fecha, AAAA-MM-DD - Día en que se realizó la medición.
- **hora**: hh:mm:ss - Hora en que se realizó la medición.
- **dsc_avenida**: Texto - Nombre de la avenida donde se mide el tráfico.
- **dsc_int_anterior**: Texto - Nombre de la vía desde donde vienen los vehículos.
- **dsc_int_siguiente**: Texto - Nombre de la vía hacia donde se dirigen los vehículos.
- **latitud**: Float - Latitud del lugar de medición.
- **longitud**: Float - Longitud del lugar de medición.
- **volumen**: Numérico - Cantidad de vehículos detectados en el carril en los últimos 5 minutos.
- **volumen_hora**: Numérico - Cantidad de vehículos detectados en el carril en la última hora.

Velocidad promedio vehicular en las principales avenidas de Montevideo

- **cod_detector:** Numérico - ID de la cámara que monitorea un carril específico para detectar vehículos.
- **id_carril:** Numérico - Número del carril monitoreado (1, 2, 3, ...).
- **fecha:** AAAA-MM-DD - Día en que se realizó la medición.
- **hora:** hh:mm:ss - Hora en que se realizó la medición.
- **dsc_avenida:** Texto - Nombre de la avenida donde se mide el tráfico.
- **dsc_int_anterior:** Texto - Nombre de la vía desde donde vienen los vehículos.
- **dsc_int_siguiente:** Texto - Nombre de la vía hacia donde se dirigen los vehículos.
- **latitud:** Float - Latitud del lugar de medición.
- **longitud:** Float - Longitud del lugar de medición.
- **velocidad_promedio:** Numérico - Promedio de las velocidades de los vehículos que circularon por el carril durante los últimos 5 minutos.

Ubicación de sensores de medición de conteo vehículos

- **dsc_avenida:** Texto - Nombre de la avenida donde se encuentra el sensor o cámara y donde se mide el tránsito.
- **dsc_int_anterior:** Texto - Nombre de la vía que forma el cruce desde donde vienen los vehículos.
- **dsc_int_siguiente:** Texto - Nombre de la vía que forma el cruce donde está el sensor. En general, el sensor se encuentra un poco antes de esta vía. El sentido de circulación será desde el cruce con **dsc_int_anterior** hacia el cruce con **dsc_int_siguiente**.
- **latitud:** Float - Coordenada que indica la latitud de la ubicación del sensor.
- **longitud:** Float - Coordenada que indica la longitud de la ubicación del sensor.

Sobre estos datos en particular, son *100 sensores* que se van cambiando de ubicación mes a mes.

Base de datos

Debido a que los datos están estrechamente relacionados y a su vez son sumamente masivos, hemos decidido utilizar una base de datos quedando de la siguiente manera.

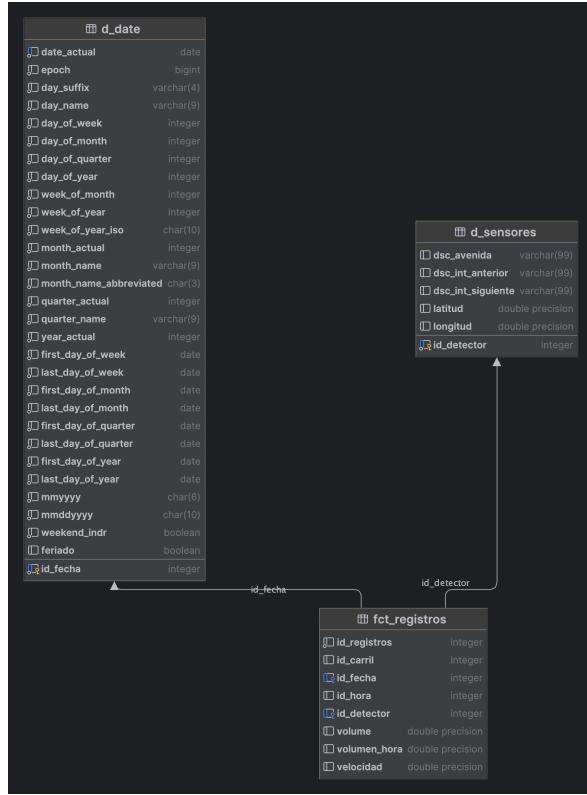


Figure 1: Diagrama de la base de datos

Nuestra tabla principal será **fct_registros**.

Tabla: fct_registros

- Cantidad de datos: 85386695.
- Variables de la tabla:
 - **id_registros**: Numérico (*Primary Key*).
 - **id_carril**: Numérico.
 - **id_fecha**: Numérico (*Foreign Key*, vinculado con **d_sensores**). La fecha de la que fue tomada el registro, tiene el formato *YYYY-MM-DD*
 - **id_hora**: Numérico. Hora en la que fue tomado el registro con formato *HHMM*.
 - **id_detector**: Numérico (*Foreign Key*, vinculado con **d_date**).
 - **volume**: Numérico. Cantidad de vehículos que pasaron en los últimos 5 minutos.
 - **volumen_hora**: Numérico. Cantidad de vehículos que pasaron en la última hora.
 - **velocidad**: Numérico. Velocidad promedio de los vehículos registrados en los últimos 5 minutos.
Unidad en km/h

Tabla: d_sensores

- Cantidad de datos: 273
- Variables de la tabla:
 - **id_detector**: Numérico (*Primary Key*).
 - **dsc_avenida**: Texto. Calle donde se encuentra el sensor.

- **dsc_int_anterior**: *Texto*. Cruce previo de la calle en **dsc_avenida**.
- **dsc_int_siguiente**: *Texto*. Cruce posterior de la calle en **dsc_avenida**. Estas dos juntas nos dirá que cada sensor se encuentra en *Avenida* entre *Anterior* y *Siguiente*.
- **latitud**: *Numérico continuo*.
- **longitud**: *Numérico continuo*. Junto a **latitud** nos indica las coordenadas geográficas del sensor.
- **barrio**: *Texto*. Esta variable fue creada a partir del paquete **geouy**

Tabla: d_date

- Cantidad de datos: 3652
- Variables de la tabla:
 - **id_fecha**: *Numérico (Primary Key)*
 - **date_actual**: *Fecha*. Secuencia de fechas desde el 01-01-2021 con formato *YYYY-MM-DD*
 - **epoch**
 - **day_suffix**: *Texto*. Fecha del día abreviado.
 - **day_name**: *Texto*. Nombre del día
 - **day_of_week**: *Numérico*. Día de la semana que indica 1 como lunes, 2 como martes, etc.
 - **day_of_month**: *Numérico*. Fecha del mes, va desde 1 hasta 31.
 - **day_of_quarter**: *Numérico*. Día del cuatrimestre.
 - **day_of_year**: *Numérico*. Día del año, del 1 al 366.
 - **week_of_month**: *Numérico*. Semana de cada mes, valores del 1 al 5.
 - **week_of_year**: *Numérico*. Semana del año, valores del 1 al 53.
 - **week_of_year_iso**: *Texto*. Variable que combina el año, la semana del año y el día de la semana.
 - **month_actual**: *Numérico*. Mes del año tomado como número, enero como 1, febrero como 2 y así sucesivamente.
 - **month_name**: *Texto*. Mes del año traducido en texto, de enero a diciembre
 - **month_name_abbreviated**: *Texto*. Mes del año en formato abreviado.
 - **quarter_actual**: *Numérico*. Indica el cuatrimestre correspondiente con números del 1 al 4.
 - **quarter_name**: *Texto*. Indica el cuatrimestre en formato de texto, primero, segundo, tercero y cuarto.
 - **year_actual**: *Numérico*. Indica el año.
 - **first_day_of_week**: *Fecha*. Indica el primer día de la semana que corresponde tal fecha.
 - **last_day_of_week**: *Fecha*. Indica el último día del rango de la semana correspondiente.
 - **first_day_of_month**: *Fecha*. Límite inferior que indica a qué mes corresponde cada fecha.
 - **last_day_of_month**: *Fecha*. Límite superior que indica a qué mes corresponde cada fecha.
 - **first_day_of_quarter**: *Fecha*. Límite inferior que indica a qué cuatrimestre corresponde cada fecha.
 - **last_day_of_quarter**: *Fecha*. Límite superior que indica a qué cuatrimestre corresponde cada fecha.
 - **first_day_of_year**: *Fecha*. Límite inferior que indica a qué año corresponde cada fecha.
 - **last_day_of_year**: *Fecha*. Límite superior que indica a qué año corresponde cada fecha.
 - **mmyyyy**: *Numérico*. Secuencia de caracteres que indica el mes y el año en formato MMYYYY
 - **mddyyyy**: *Numérico*. Secuencia de caracteres que indica el mes, la fecha y el año en formato MMDDYYYY.
 - **weekend_indr**: *Lógico*. TRUE si la fecha tiene como día de la semana sábado o domingo, FALSE en caso contrario.
 - **feriado**: *Lógico*. TRUE si la fecha correspondiente coincide con días feriados en Uruguay, FALSE en caso contrario.

La descripción de los datos podrían haberlo hecho con una sola tabla, la explicación es confusa y muy redundante. Cuando explican la diferencia entre los datos originales y los que se almacenaron en la base de datos no explican la diferencia que existe y por qué se decidió hacerlo así, no era obligatorio pero la descripción así no tiene sentido, menos, es más.

Cuando describen la base de datos, hacer una tabla era mejor, la captura de pantalla era la misma que les adjuntó Mauro por correo, no está referenciada ni enumerada.

Análisis exploratorio

No es homogenea en que sentido, en que se basan?

En nuestro proyecto tenemos datos que tienen una dimension geo-espacial, por lo que es importante tener en cuenta que la información que tenemos no es homogenea. Tambien es importante tener en cuenta que la información que tenemos es de un periodo de tiempo acotado.

Dicho esto, para empezar, me parecio adecuado comprobar la integridad de los datos, es decir, ver si tenemos datos faltantes o datos que no tienen sentido.

Son tres y escribe solo uno?

Datos faltantes y nulos

```
##      atributo cant_total cant_null cant_0 porc_null   porc_0
## 1    velocidad     85386695          0 8873659 0.000000 10.39232
## 2 volumen_hora    85386695          0 8873659 0.000000 10.39232
## 3      volume     85386695          0 8873659 0.000000 10.39232
## 4     id_fecha    85386695          0          0 0.000000 0.00000
## 5   id_detector    85386695        2274          0 0.002663 0.00000
```

En 2274 datos se perdio la informacion de la ubicacion del sensor, por lo que no sabemos de que calle se trata. En el 10.39232% de los datos se detecto velocidad 0 y en el mismo porcentaje se detecto volumen 0.

Se descubrio que la cantidad de datos que tienen los tres campos en 0 es 8873659 registros, lo que representa el 100% de los datos nulos.

De los 8873659 registros que tienen los tres campos en 0, 264 no tienen id_detector y 8873395 si tienen id_detector.

No queda claro si los datos que tienen los tres campos en 0 son datos que representan que no paso ningun vehiculo por el sensor o si son datos que no se pudieron obtener.

Sobre los datos faltantes de la ubicacion del sensor, son datos que no se pueden recuperar, por lo que se tendran que descartar.

Se quiso averiguar en que fecha se perdio la informacion de la ubicacion del sensor, para ver si se podia recuperar la informacion de otra forma, pero no se pudo.

```
##   id_fecha cant_total cant_null porc_null
## 1 20210724     131083      259  0.197585
## 2 20210725     131607      287  0.218074
## 3 20210726     132574      288  0.217237
## 4 20210727     132635      288  0.217137
## 5 20210728     131988      288  0.218202
## 6 20210729     130631      288  0.220468
## 7 20210730     130224      288  0.221157
## 8 20210731     129410      288  0.222548

##      id_registros      id_carril      id_fecha      id_hora
## Min.   :91804143   Min.   :2   Min.   :20210724   Min.   : 0
## 1st Qu.:91954682   1st Qu.:2   1st Qu.:20210726   1st Qu.: 610
## Median :92116584   Median :2   Median :20210728   Median :1205
## Mean   :92111331   Mean   :2   Mean   :20210728   Mean   :1192
## 3rd Qu.:92265955   3rd Qu.:2   3rd Qu.:20210730   3rd Qu.:1800
## Max.   :92403719   Max.   :2   Max.   :20210731   Max.   :2355
##      id_detector      volume      volumen_hora      velocidad
## Mode:logical      Min.   : 0.000   Min.   : 0.00   Min.   : 0.00
```

```

## NA's:2274      1st Qu.: 2.000    1st Qu.: 24.00    1st Qu.:37.00
##                           Median : 4.000    Median : 48.00    Median :43.00
##                           Mean   : 4.978    Mean   : 59.74    Mean   :38.39
##                           3rd Qu.: 7.000    3rd Qu.: 84.00    3rd Qu.:47.00
##                           Max.   :37.000    Max.   :444.00    Max.   :90.00
##       fecha           hora
##   Min.   :2021-07-24    Min.   : 0
##   1st Qu.:2021-07-26    1st Qu.: 610
##   Median :2021-07-28    Median :1205
##   Mean   :2021-07-27    Mean   :1192
##   3rd Qu.:2021-07-30    3rd Qu.:1800
##   Max.   :2021-07-31    Max.   :2355

```

Se puede ver que los datos faltantes de la ubicacion del sensor van desde el 24/07/2021 hasta el 31/07/2021. Tambien se puede ver que todos los datos faltantes son del carril 2. Quizá se podrían recuperar los datos de la ubicación del sensor revisando los datos originales pero es irrelevante ya que son pocos datos y no afectan al análisis.

Datos faltantes en la base de datos

	##	count	month_actual	year_actual
## 1	3830149		1	2022
## 2	3898585		1	2023
## 3	3714726		2	2021
## 4	3374893		2	2022
## 5	3455550		2	2023
## 6	3908628		3	2021
## 7	3746326		3	2022
## 8	3879877		3	2023
## 9	3759551		4	2021
## 10	3620380		4	2022
## 11	3930695		5	2021
## 12	2701766		5	2023
## 13	3873954		6	2021
## 14	3731149		6	2022
## 15	4078328		7	2021
## 16	3679998		7	2022
## 17	4032694		8	2021
## 18	3857623		9	2021
## 19	2501798		9	2022
## 20	3748052		10	2021
## 21	4072685		10	2022
## 22	3979504		11	2021
## 23	4009784		12	2021

Información presentada con las salidas de R así nomás no está bien, además que mucha información es constante, podría escribirse en el texto. Con respecto a lo que comentan sobre la información con volumen y velocidad cero, podrían haber incluido un gráfico de barras y generar un mayor impacto visual.

En el summary hay información que no debería estar allí como las claves subrogadas. El summary debe de presentarse de una forma más prolífica y que tenga un sentido.

Haciendo una observación de los meses cubiertos por los datos, nos hemos dado cuenta que faltan algunos meses en la base de datos.

Ubicacion de los sensores

Los datos que tenemos son de 100 sensores ubicados todos en Montevideo y estos van cambiando de ubicacion cada mes. Por lo que el primer paso es ver cuantas ubicaciones distintas tenemos y cuantos sensores hay en cada ubicacion.

Para mostrarlo, hemos decidido utilizar un mapa de Montevideo con los barrios y mostramos la cantidad de sensores ubicados en el. En el mapa se puede ver que los sensores estan ubicados en 42 de los 62 barrios de Montevideo. Los barrios que tienen sensores son 42 sobre 62 siendo los barrios de Buceo, Centro, Pocitos y Unión con mas de 20 sensores.

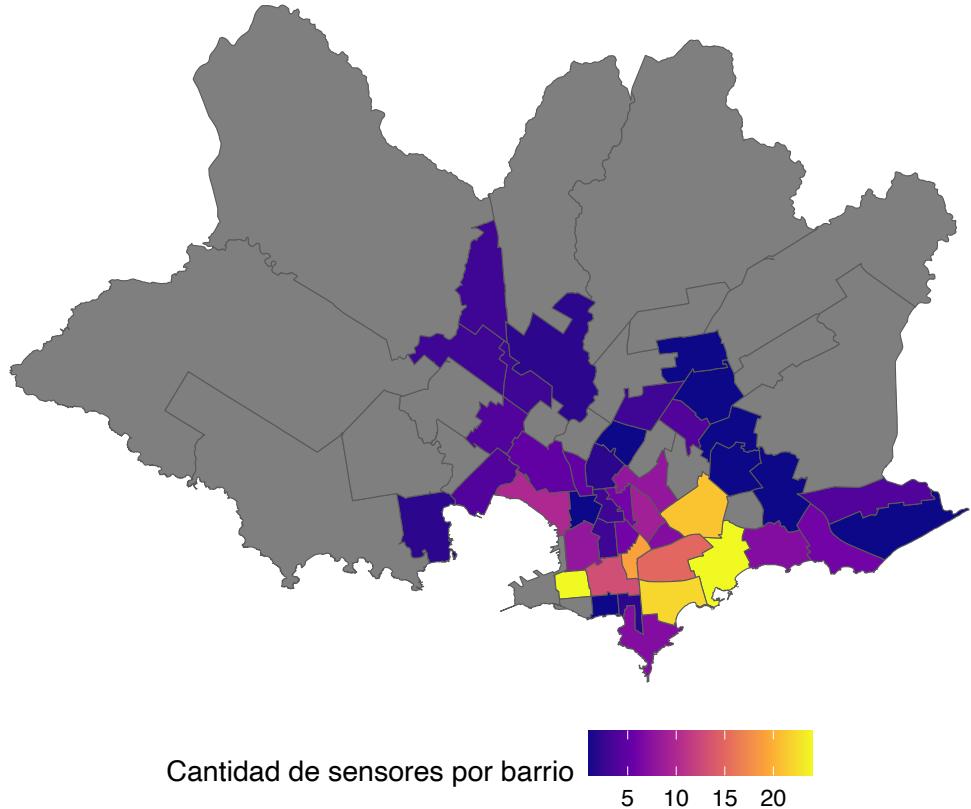


Figure 2: Mapa de Montevideo con cantidad de sensores por barrio.

Ahora quiero mostrar la cantidad de datos que tenemos por ubicacion, para ver si hay alguna ubicacion en particular que tenga mas o menos datos que las demás.

Gráfico apropiado, ¿pero hay algo que les llame la atención?. Luego en el siguiente pasaje con la cantidad de sensores por barrio, no explicaron de donde y como obtuvieron la densidad de sensores por metro cuadrado. Vimos que utilizaron los datos de RLadies pero es relevante explicarlo ya que esta fuente de datos no la incluyen cuando hacen la descripción.

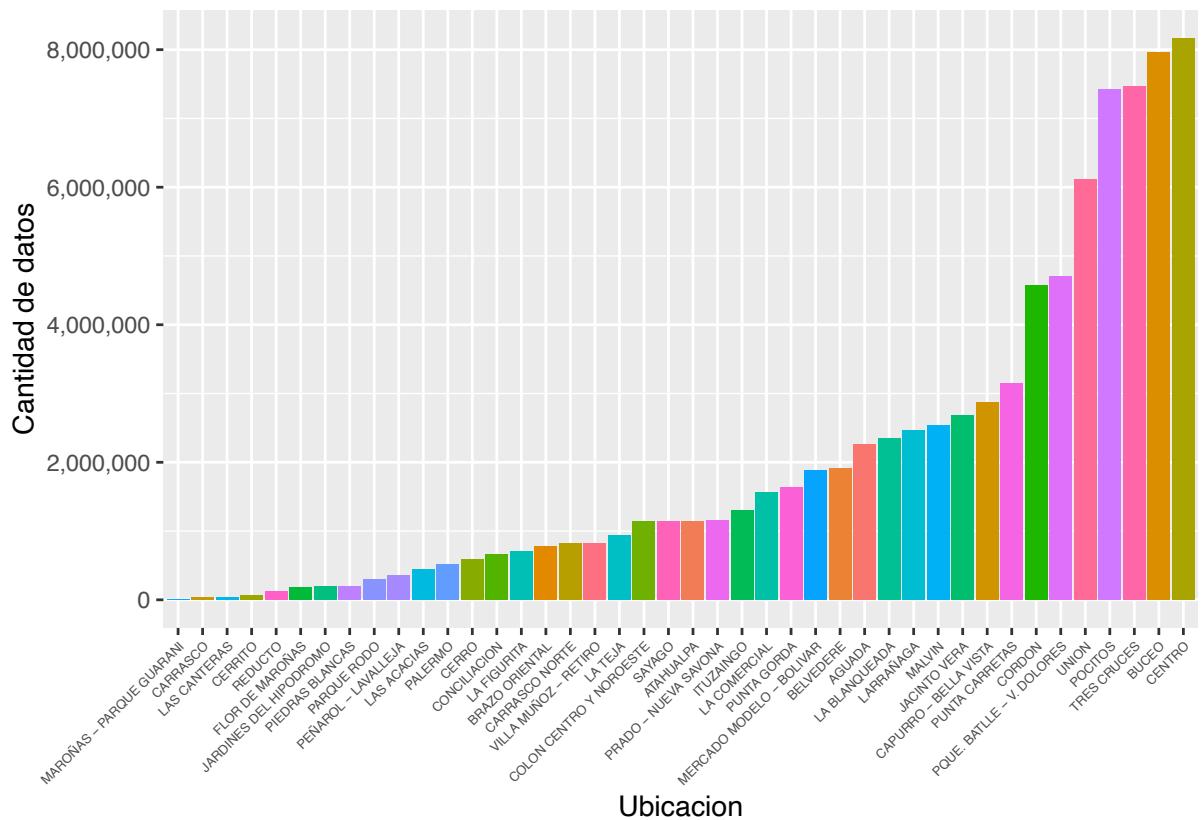


Figure 3: Cantidad de datos por ubicacion

Se puede observar que la cantidad de datos por ubicacion no es para nada homogenea. Los barrios con mayor cantidad de datos aportados al dataset son Union, Pocitos, Tres Cruces, Buceo y Centro. Por otro lado Maroñas, Carrasco, Las Canteras y Cerrito son los que menos datos aportan.

Ahora me interesaría saber cuales son los barrios mejores representados en el dataset, es decir, cuales son los barrios que tienen mas datos por metro cuadrado.

primero calculo el area de cada barrio y luego calculo la cantidad de datos por metro cuadrado.

Texto entreverado, muchas faltas y frases sueltas. Esto es relevante y tienen que mejorar en futuros trabajos

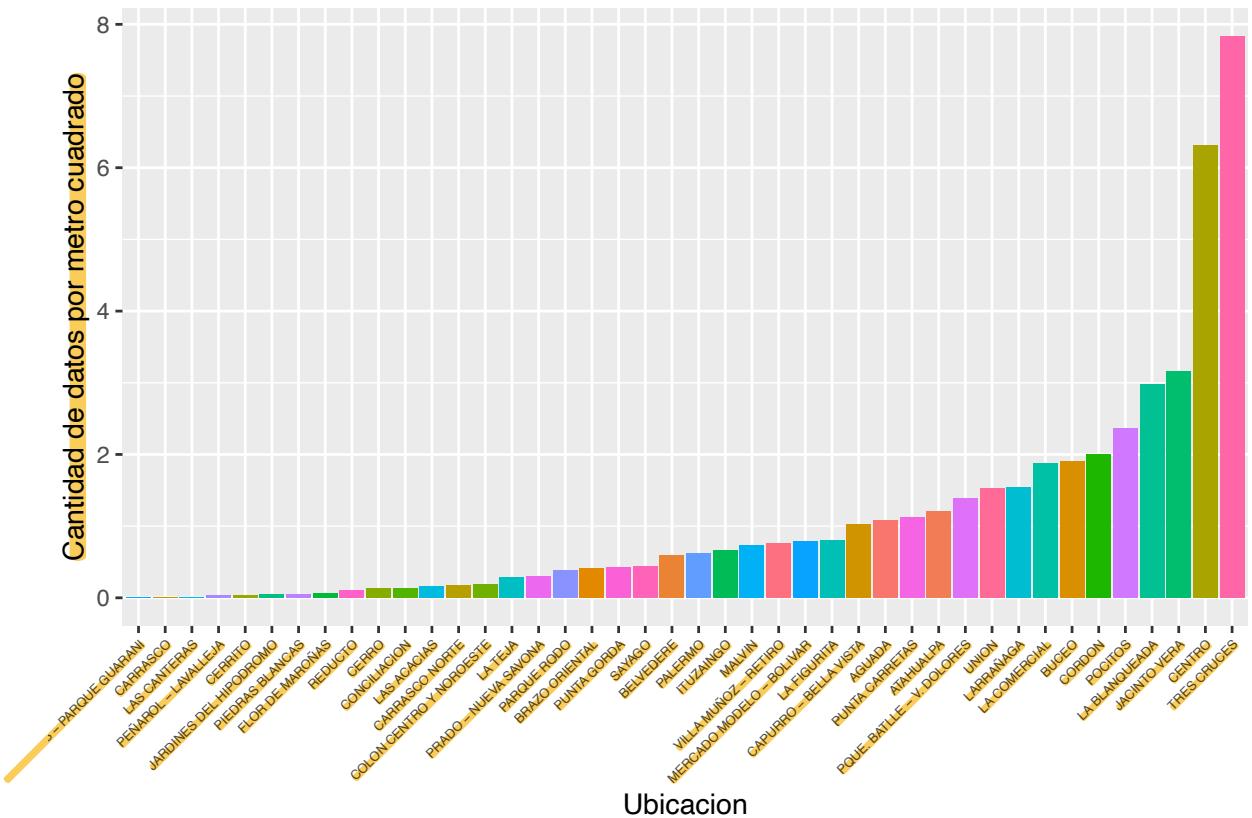


Figure 4: Cantidad de datos por ubicacion ponderado por area

Se puede observar que los barrios con mayor cantidad de datos por metro cuadrado son Pocitos, La Blanqueada, Jancito Vera, Centro y Tres Cruces. Por otro lado Maroñas, Carrasco, Las Canteras, Peñarol y Cerrito son los que menos datos aportan por metro cuadrado.

Principales variables

Las variables que se van a analizar son las siguientes:

- **volume**: Numérico. Cantidad de vehiculos que pasaron en los últimos 5 minutos.
- **volumen hora**: Numérico. Cantidad de vehiculos que pasaron en la ultima hora.
- **velocidad**: Numérico. Velocidad promedio de los vehiculos registrados en los ultimos 5 minutos. Unidad en km/h

Velocidad

Resumen de la variable velocidad

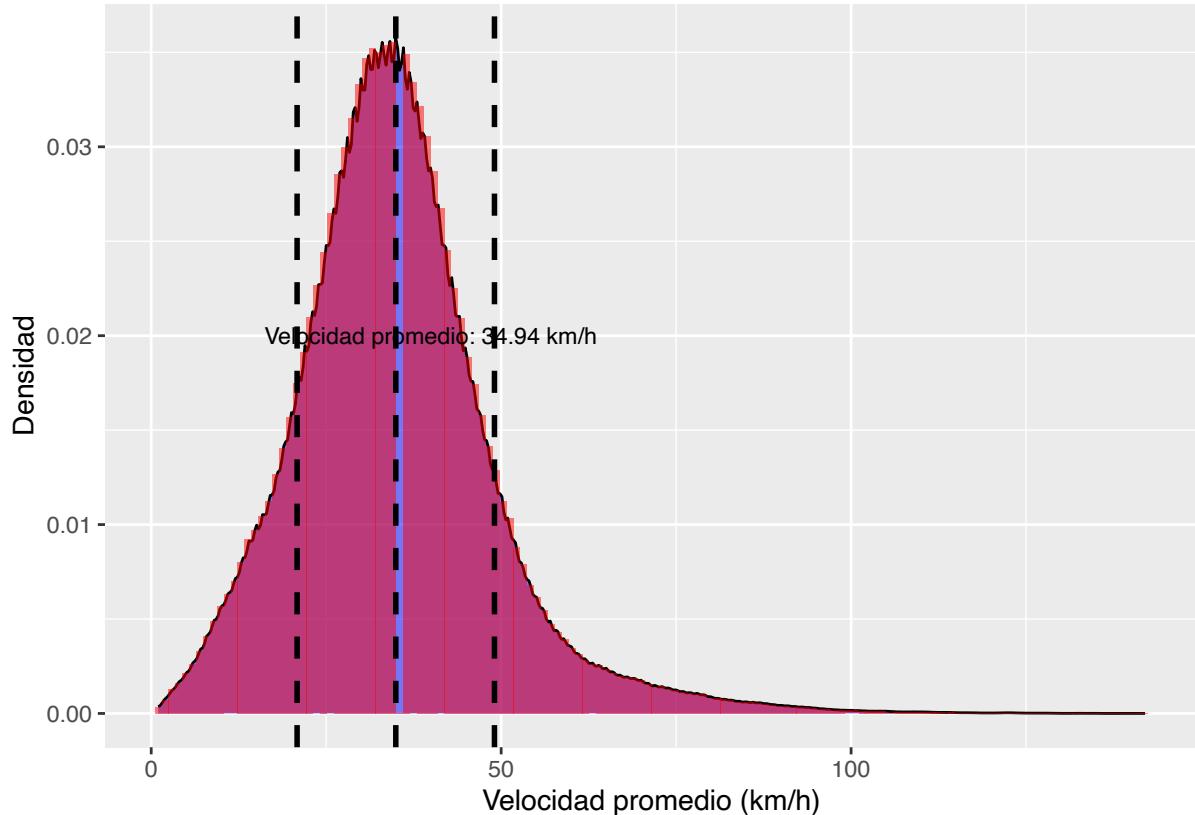
No se entiende está muy desprolijio

```
minimo primer_cuartil mediana tercer_cuartil maximo promedio desvio 1 0 22 32 41 144 31.31286 17.07865
```

Ahora veamos la distribucion de la velocidad registrada Para mejor visualizacion, voy a dejar de lado los datos donde la velocidad es 0

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## Please use 'linewidth' instead.
## This warning is displayed once every 8 hours.
```

```
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Hubiese aportando un análisis con mayor desagregación, hora, día....

Figure 5: Distribucion de velocidad promedio

no podés asegurar que su distribución es normal

Se puede observar que la distribucion de la velocidad es normal, con una media de 31.31 km/h y un desvio de 17.08 km/h. El 68% de los datos se encuentran entre 14.23 km/h y 48.39 km/h.

Volumen

Resumen de la variable volumen

minimo, maximo, promedio, desvio, cuartiles

```
##   minimo cuartil_1 mediana cuartil_3 maximo promedio   desvio
## 1      0          3         11        26       659 17.27588 19.18341
```

Se ve que la cantidad minima de vehiculos registrados en 5 minutos es 0 y la maxima es 659. El promedio de volumen es 17.28 vehiculos y el desvio estandar es 19.18 vehiculos.

Ahora veamos la distribucion del volumen registrado Para mejor visualizacion, voy a dejar de lado los datos donde el volumen es 0

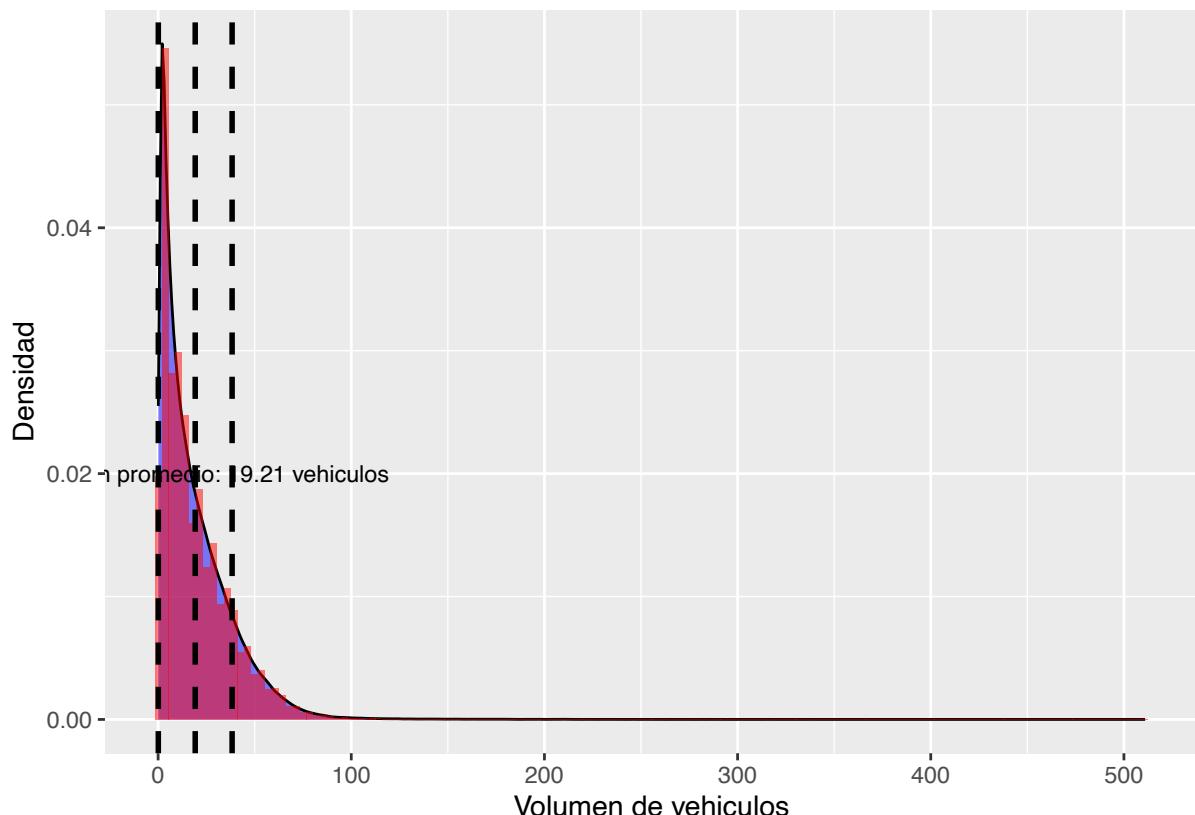


Figure 6: Distribucion de volumen

Se puede observar que la mayoría de los valores son menores a 100, en particular el 75% de los datos son menores a 26 vehículos.

Preguntas de Investigacion

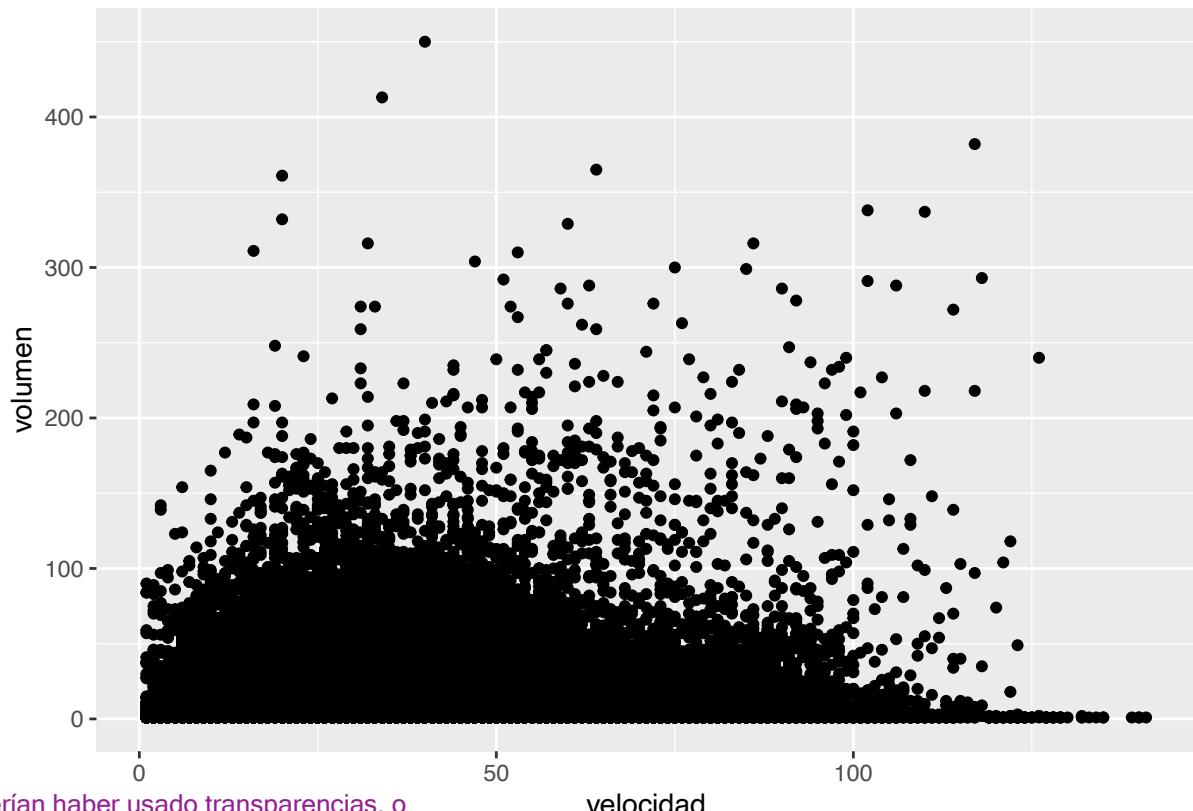
Las preguntas que dirigirán este análisis son las siguientes:

1. ¿Existe alguna correlación entre el volumen y la velocidad?.
2. ¿Cuáles son las calles con los mayores promedios de velocidad en Montevideo? ¿Con qué frecuencia se cometen excesos de velocidad?.
3. ¿Cómo va variando el volumen y velocidad medidos a través del TIEMPO?.

1. ¿Existe alguna correlación entre el volumen y la velocidad?

Haremos un grafico de puntos para visualizarlo. Para los datos usaremos una muestra aleatoria de toda la base de datos

Mencionar el problema de sobreploteo y que decidieron muestrear datos para ver algo. El muestreo debería ser estrechificado de alguna forma o proporcional al tamaño para mantener la estructura de los datos similar



Deberían haber usado transparencias, o histograma en dos dimensiones

Figure 7: Grafico de puntos de velocidad y volumen

Definitivamente **no hay una relacion lineal** entre velocidad y volumen

```
##           velocidad      volumen
## velocidad 1.00000000 -0.02730281
## volumen   -0.02730281  1.00000000
```

La correlacion dio -0.03, lo que indica que no hay una correlacion lineal entre las variables.

2. ¿Cuáles son las calles con los mayores promedios de velocidad en Montevideo? ¿Con qué frecuencia se cometen excesos de velocidad?

Pasemos a investigar las calles con mayor promedio de velocidad.

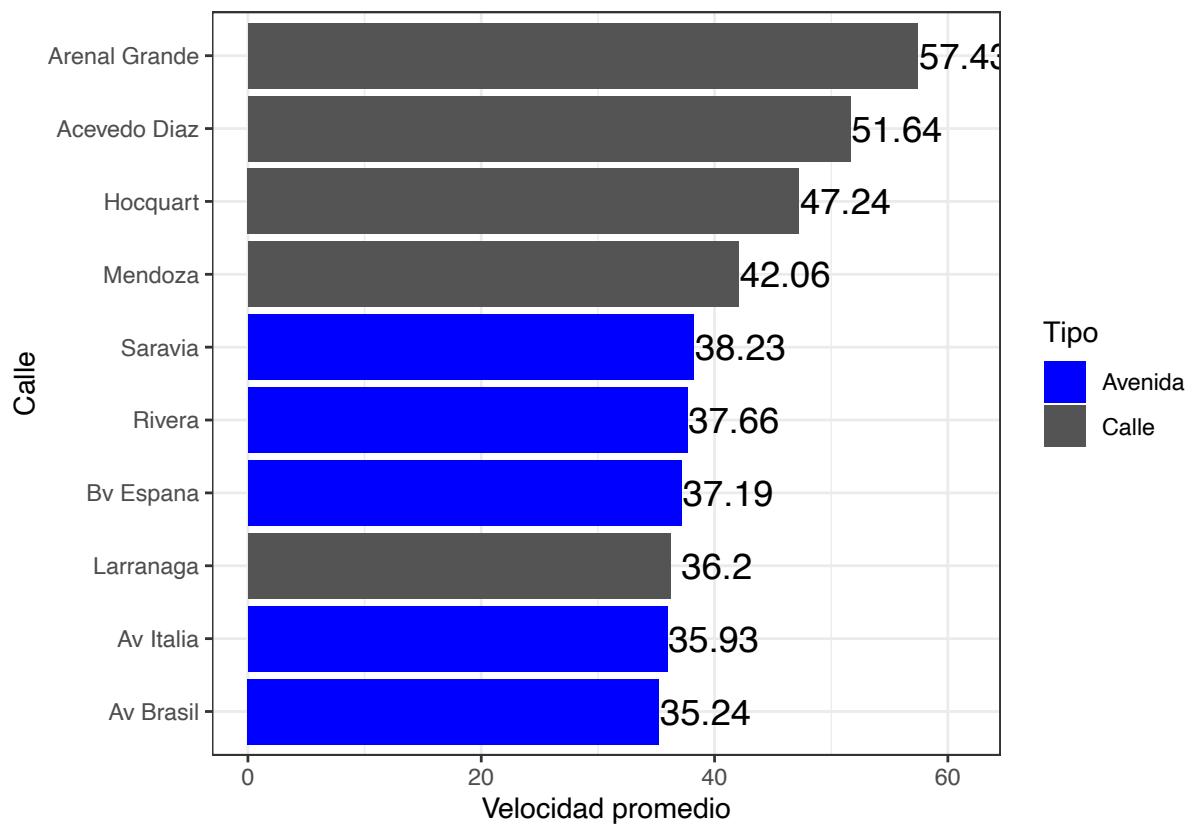


Figure 8: Calles con mayor promedio de velocidad

El siguiente gráfico nos muestra que las 3 calles con mas velocidad en promedio superan el máximo de 45km/h siendo este la velocidad máxima de circulación reglamentaria. La avenida más rápida en promedio no alcanza el máximo de velocidad permitido

En promedio de velocidad circulacion los conductores son prudentes, aun asi vemos con que frecuencia se comenten excesos de velocidad.

No todas estas calles tiene límite de 45, por ejemplo Av. Italia

Deberían usar referencias cruzadas para referir a los gráficos en el texto

No se entiende que es este gráfico

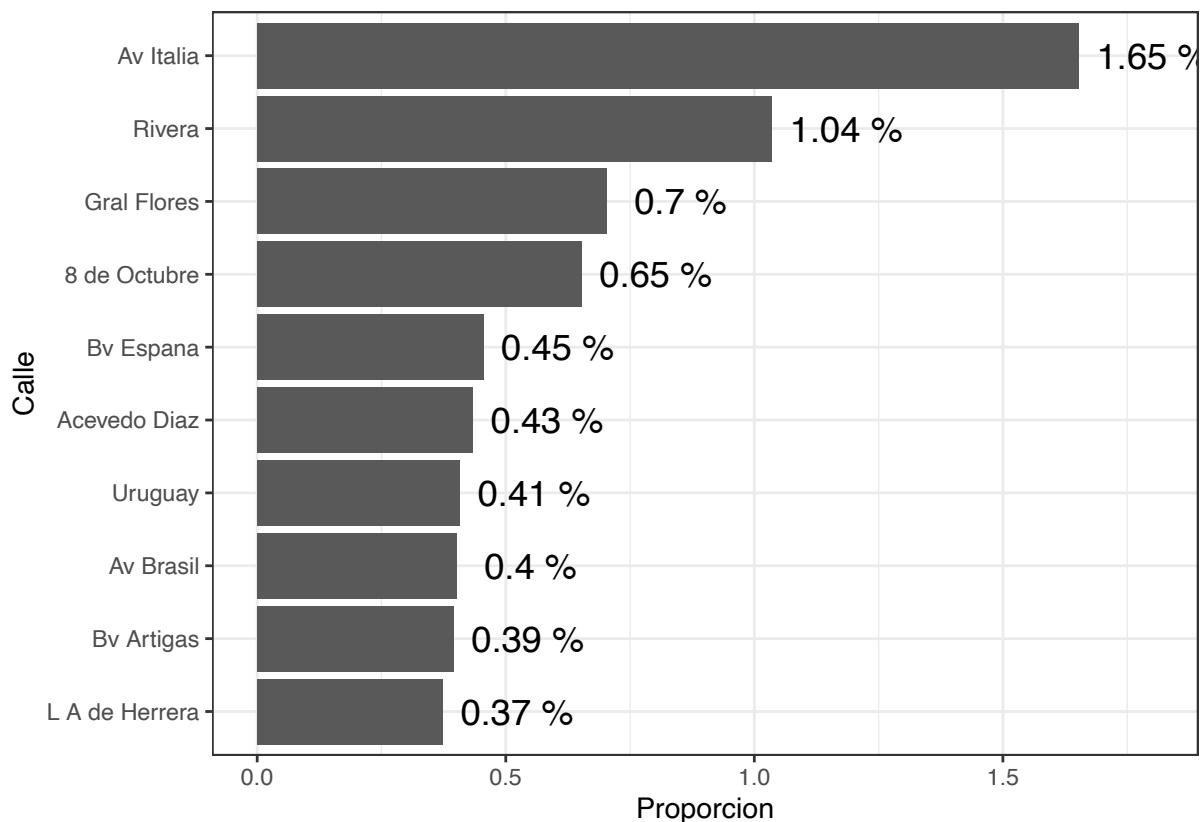


Figure 9: Proporcion de excesos de velocidad por calle

Si bien es una proporción de valor bajo tengamos en cuenta que estamos tomando el 1% de mas de 85 millones de registros, es decir que hay mas de 850 mil infracciones de velocidad solo en Avenida Rivera, en Avenida Italia hay mucha mas diferencia superando el millón cuatrocientos en casi 2 años y medio.

3. ¿Cómo va variando el volumen y velocidad medidos a traves de el tiempo?

Volumen

Primero observemos como va variado volumen y volumen hora

Variacion por hora del dia

```
## Coordinate system already present. Adding new coordinate system, which will
## replace the existing one.
```

Muy apropiada visualización para el problema, y tratando de desagregar más la información, sin embargo no comentan demasiado sobre el mismo y se ven varias cosas interesantes

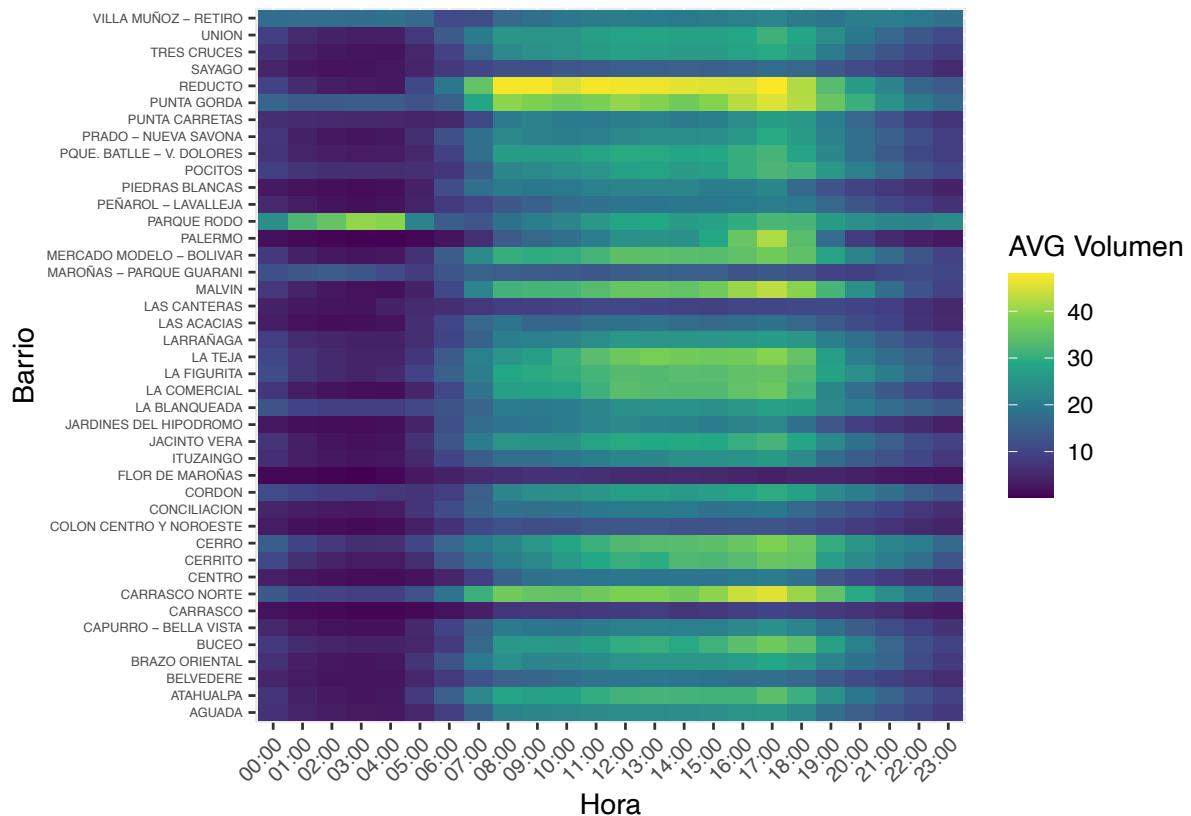
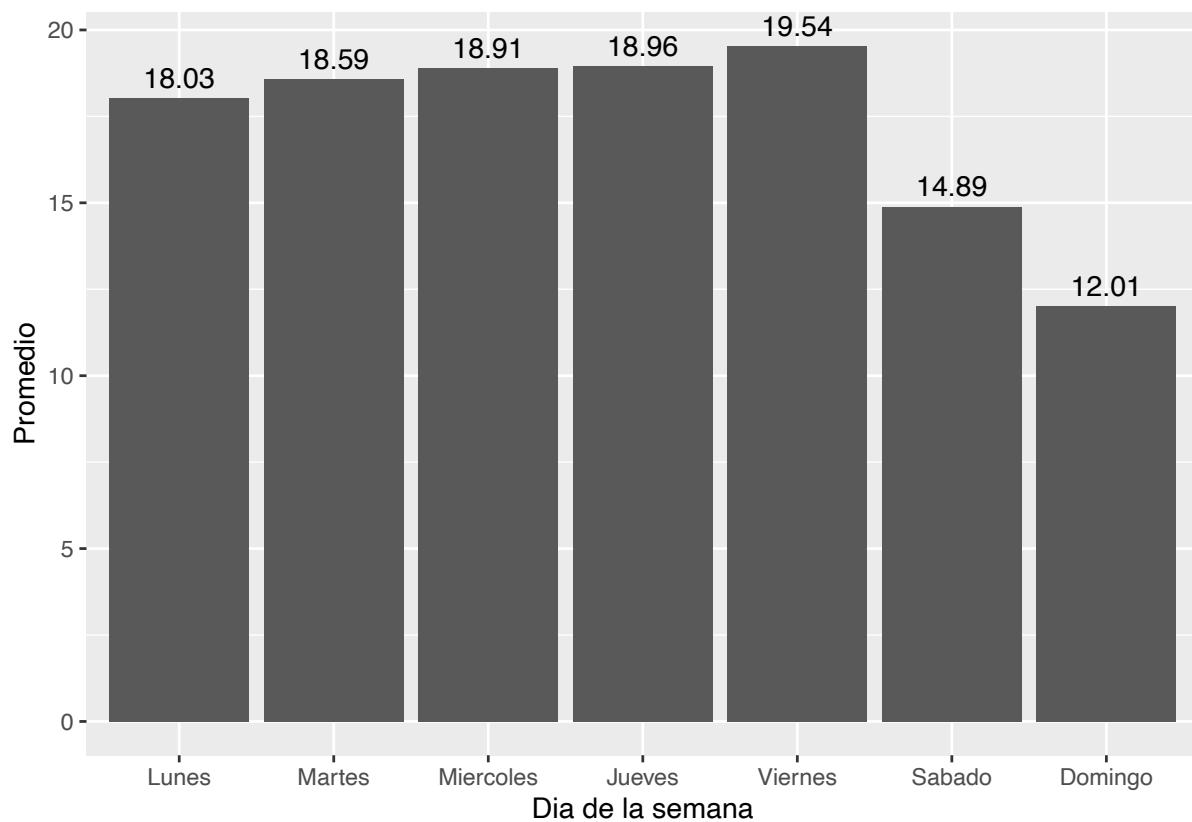


Figure 10: Mapa de calor de volumen maxima por barrio por rango de hora

En Parque Rodo hay valores altos durante la noche, que lo diferencia del resto de los barrios.

Variacion del volumen por dia de la semana Se puede notar un volumen claramente más alto los días de semana, yendo en aumento de lunes a viernes, hasta bajar el promedio los fines de semana. El día de la semana con más volumen de tránsito es el viernes, y el de menor volumen es el domingo.



Falta caption de la figura

Se puede notar un comportamiento casi idéntico en la cantidad de observaciones del volumen, según el día de la semana. La mayoría de los datos se concentra cuando el volumen está entre 0 y 10, la mayoría de los días. Los fines de semana (En especial el domingo) se puede observar un comportamiento distinto, ya que cuando el volumen está entre 0 y 10, la cantidad de observaciones es mucho mayor, y no hay tanta concentración de datos cuando el volumen aumenta.

no queda claro en que unidades está, son 18 autos cada 5 minutos en promedio el lunes?
 Deben interpretar en términos del problema no tienen título y la interpretación es incompleta.
 Los gráficos deben ser autocontenidos

"Se puede notar un comportamiento casi idéntico en la cantidad de observaciones del volumen, según el día de la semana.". Entendí lo que quisieron decir porque lo escuché en el oral. Solo con la redacción me costaría entender si estaban contando observaciones de volumen era más simple "distribución del volumen por día".

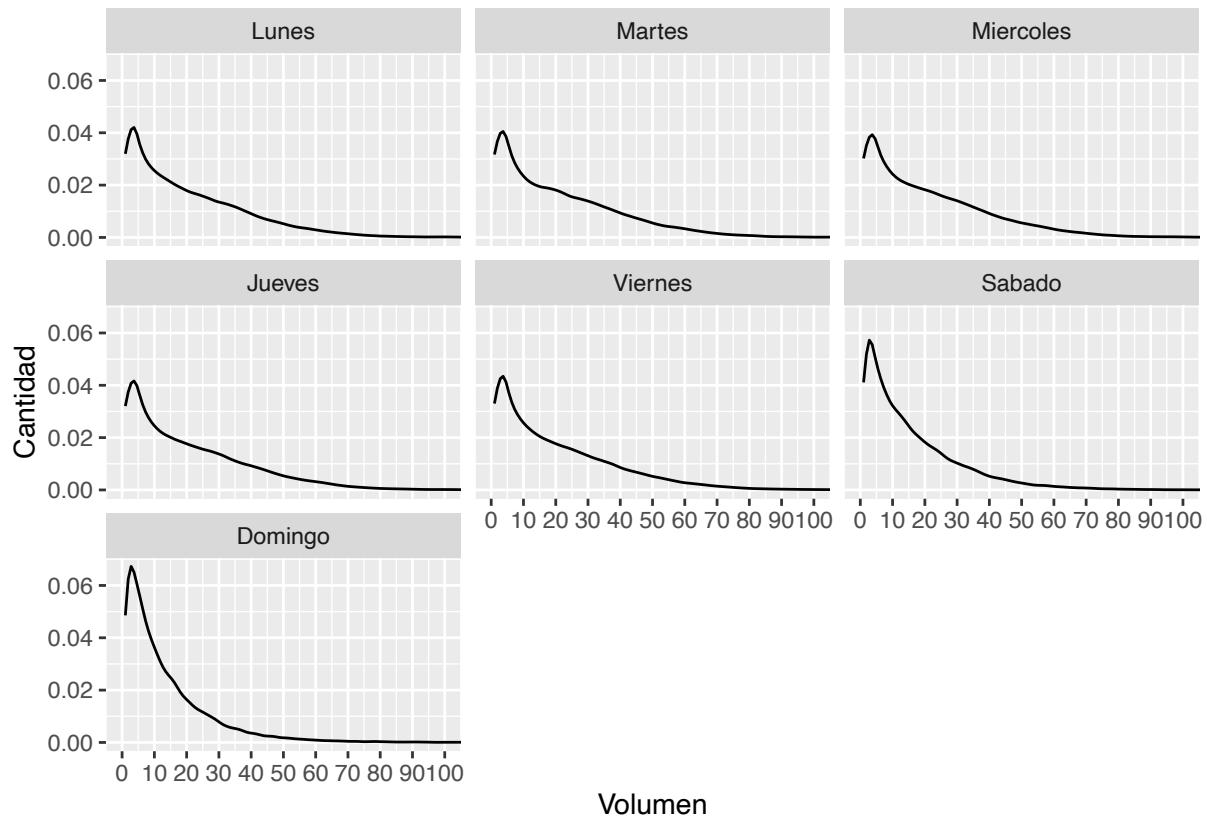
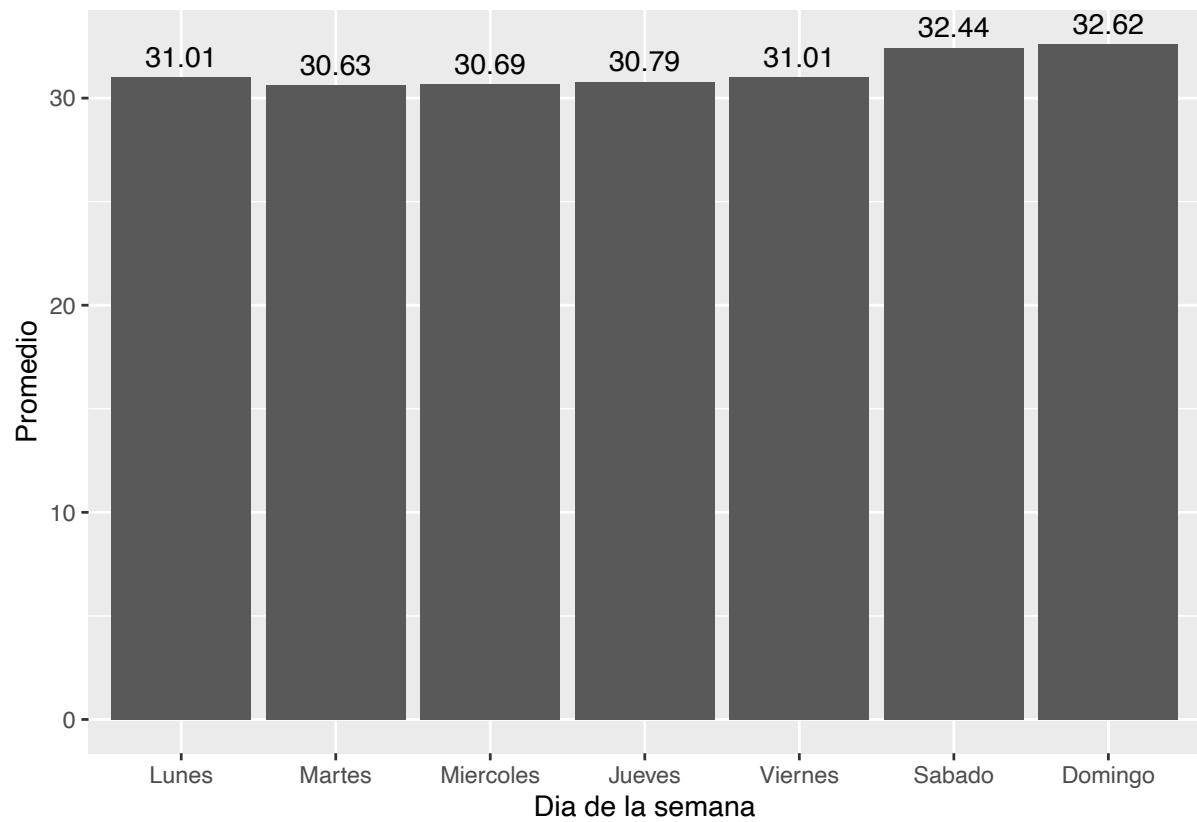


Figure 11: Densidad de Volumen

Variacion de la velocidad por dia de la semana Podemos observar que el promedio de velocidad es casi el mismo para cada día de la semana, a excepción de los fines de semana, donde la velocidad aumenta ligeramente. Esto concuerda con la disminución del volumen los fines de semana, visto anteriormente.



Viendo la densidad de la velocidad parece no variar a lo largo de los días de la semana. Alcanzan su valor máximo alrededor de los 35km/h.

titulo figura! No aporta
mucho podrían solo
comentr en el texto

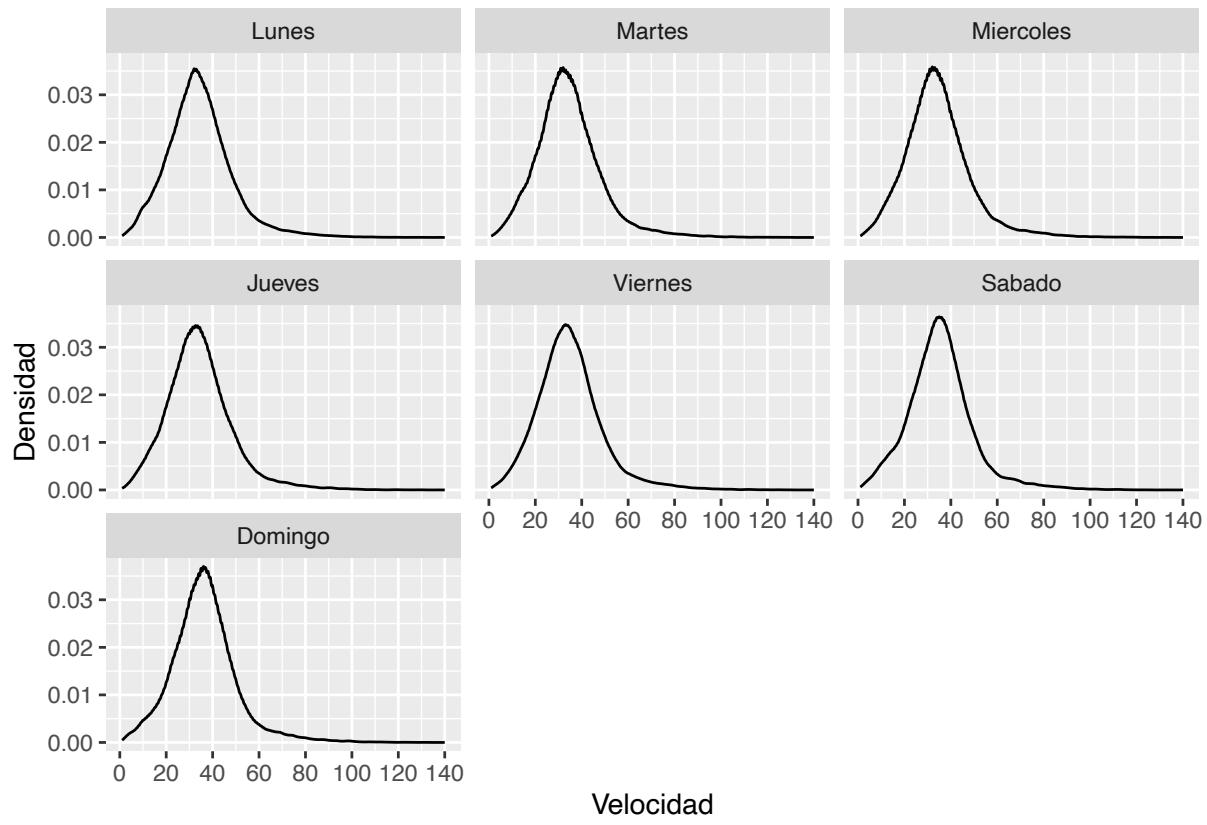


Figure 12: Densidad de Velocidad

Dentro de los días de la semana hemos observado diferentes comportamientos de la velocidad durante los días sábado y domingo (fin de semana), en particular observemos las velocidades máximas registradas de lunes a viernes y durante el fin de semana.

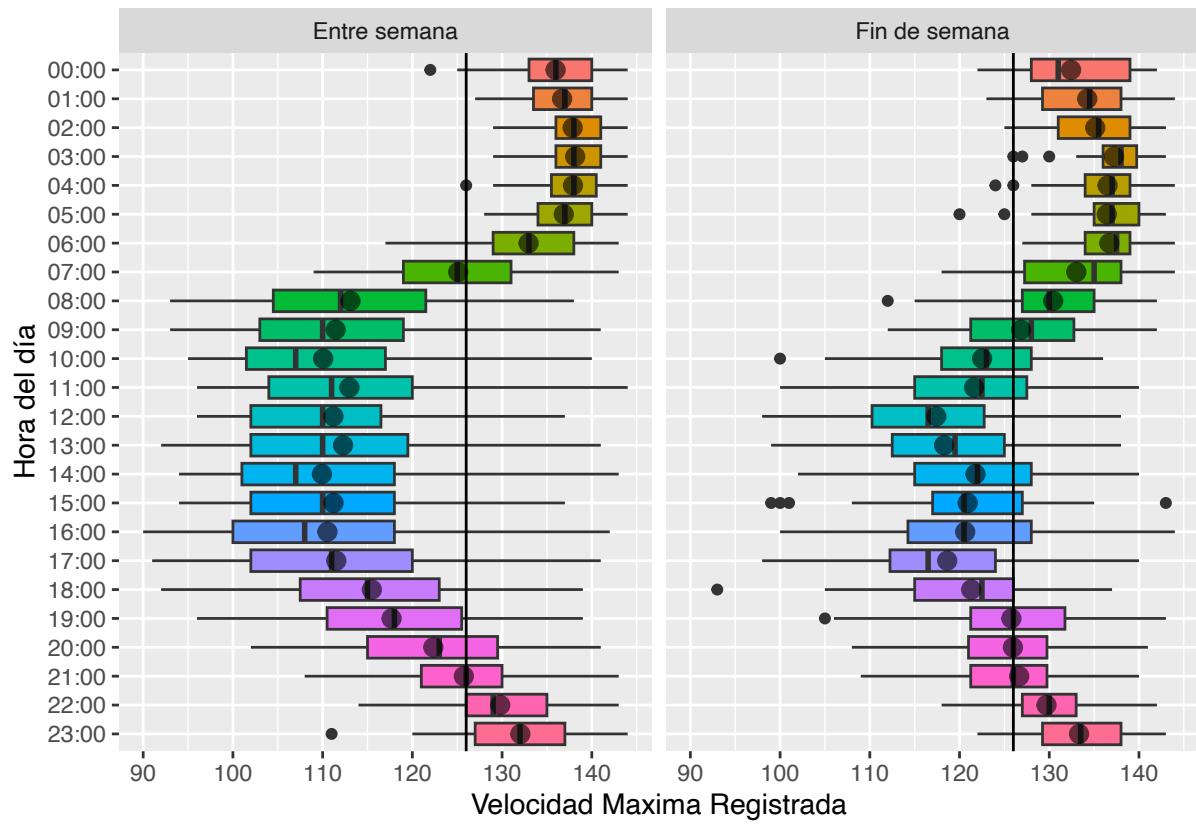


Figure 13: Distribucion maxima velocidad por hora

Este gráfico nos revela cosas interesantes, ya que podemos notar que los fines de semana, los máximos de velocidad registrados son mayores que los días de semana. De manera contraria, los máximos de velocidad con valores bajos son más comunes los días de semana. Esto también está muy relacionado a lo visto anteriormente sobre el volumen, ya que los fines de semana el volumen es menor, y eso permite facilidad a alcanzar picos de velocidad mayores. Como los días de semana el volumen de tráfico es mayor, es esperable que los máximos de velocidad no escalen demasiado.

Muy buen intento de desagregar los datos y analizar otro resumen que no sea el promedio.
Interesante el análisis del velocidad máxima que hicieron.

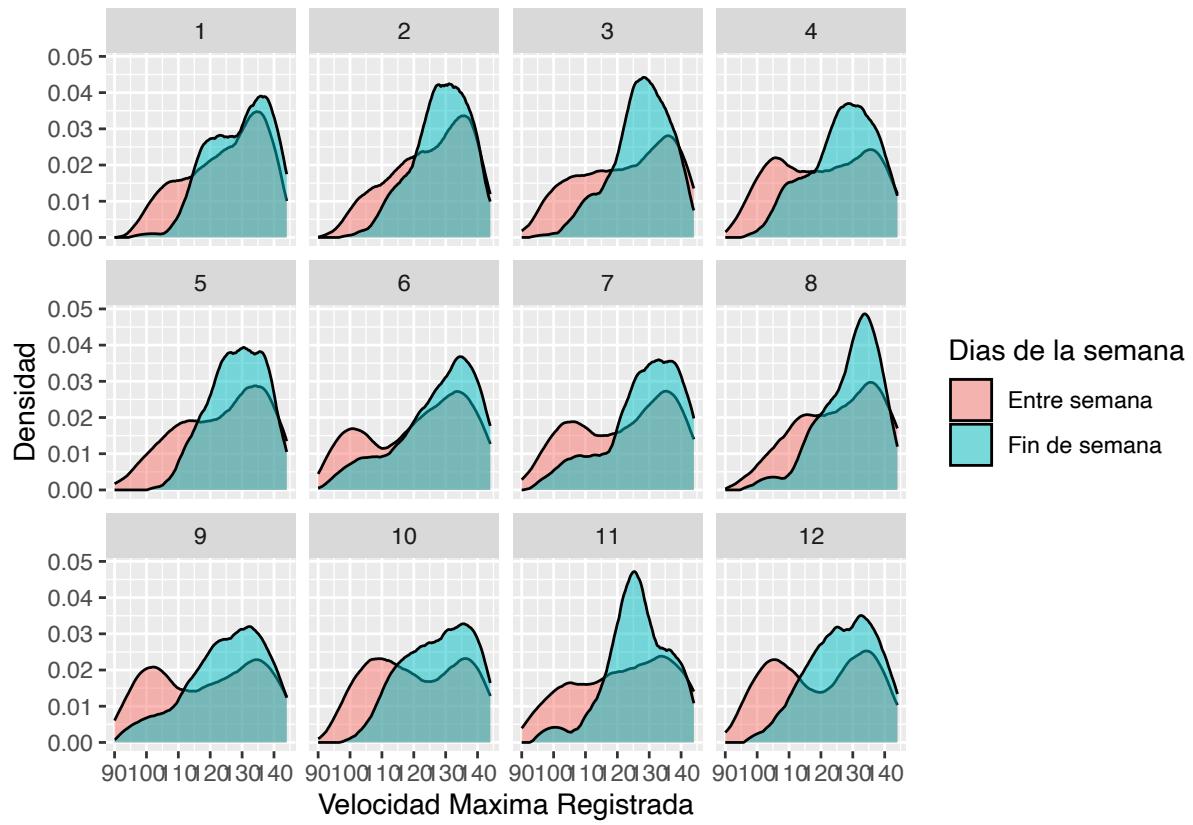


Figure 14: Densidad de la velocidad por meses

Luego, con respecto a los meses, se puede ver que los dos meses en los cuales la densidad de la velocidad máxima es más similar los días de semana y los fines de semana son enero y febrero, y esto puede ser causado por las vacaciones de verano.

Variacion por año Si entramos en cada año vemos como crece la densidad de velocidad máxima registrada de lunes a viernes.

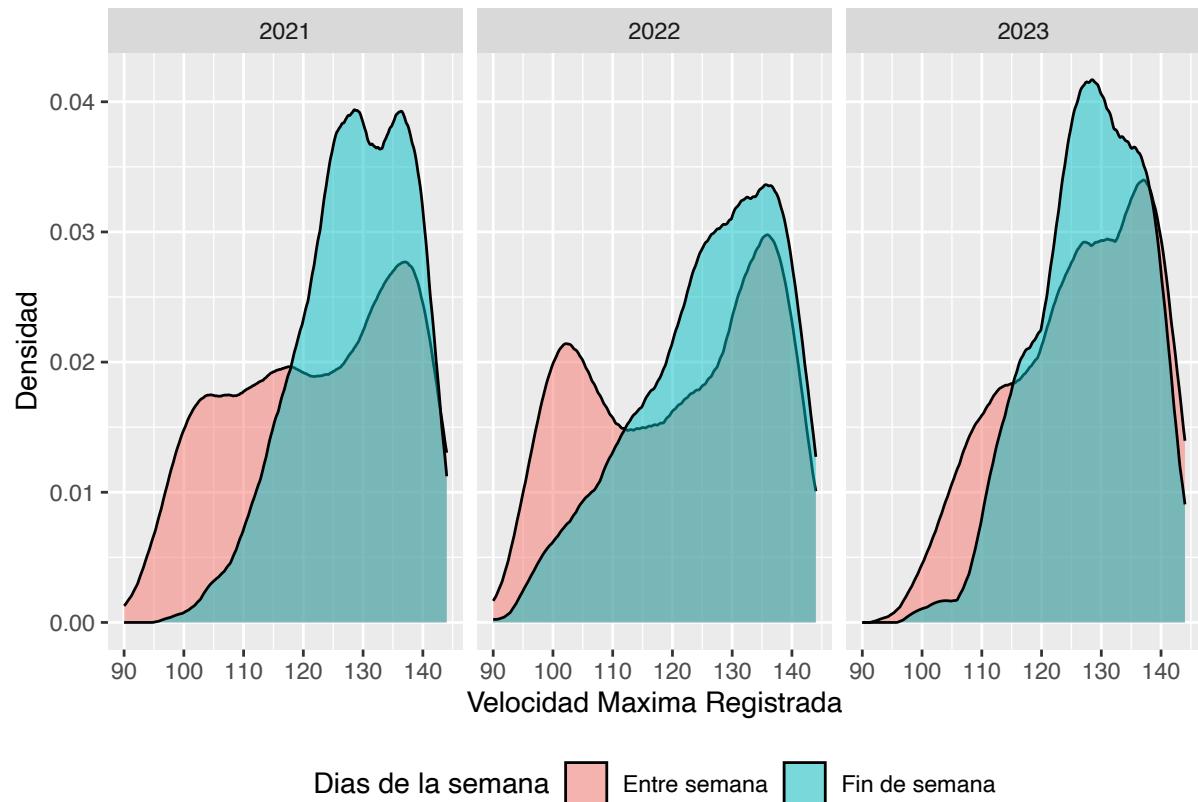
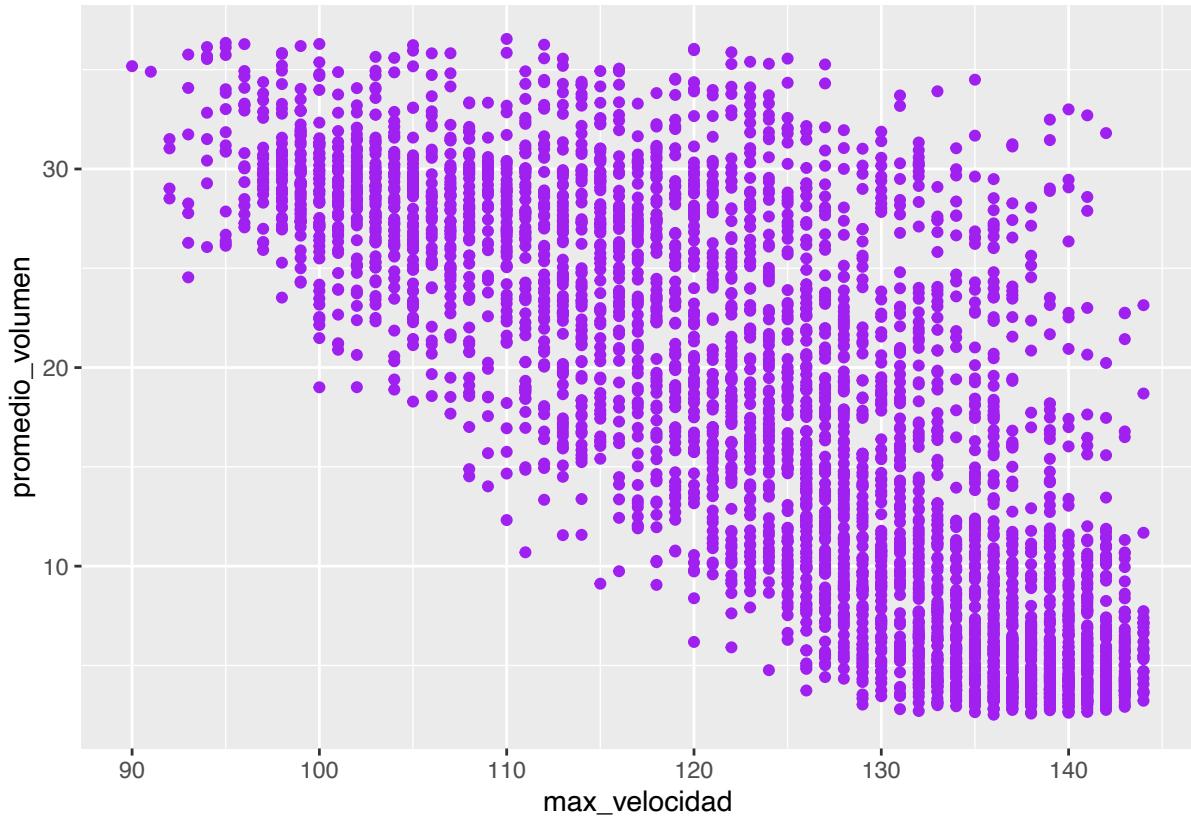


Figure 15: 2023 no esta completo

En 2021 y 2022, los días de semana hubo una densidad de velocidad máxima muy similar, aunque los fines de semana hubo mayor densidad en valores altos de velocidad en 2021. En 2023 podemos ver una densidad mayor los días de semana, con menos diferencia de los días de semana como los años anteriores. Cabe aclarar que los datos de 2023 solo abarcan hasta mayo. Esto puede dar una explicación de la diferencia menor entre fin de semana y día de semana, ya que anteriormente vimos que los primeros meses del año muestran un comportamiento similar.

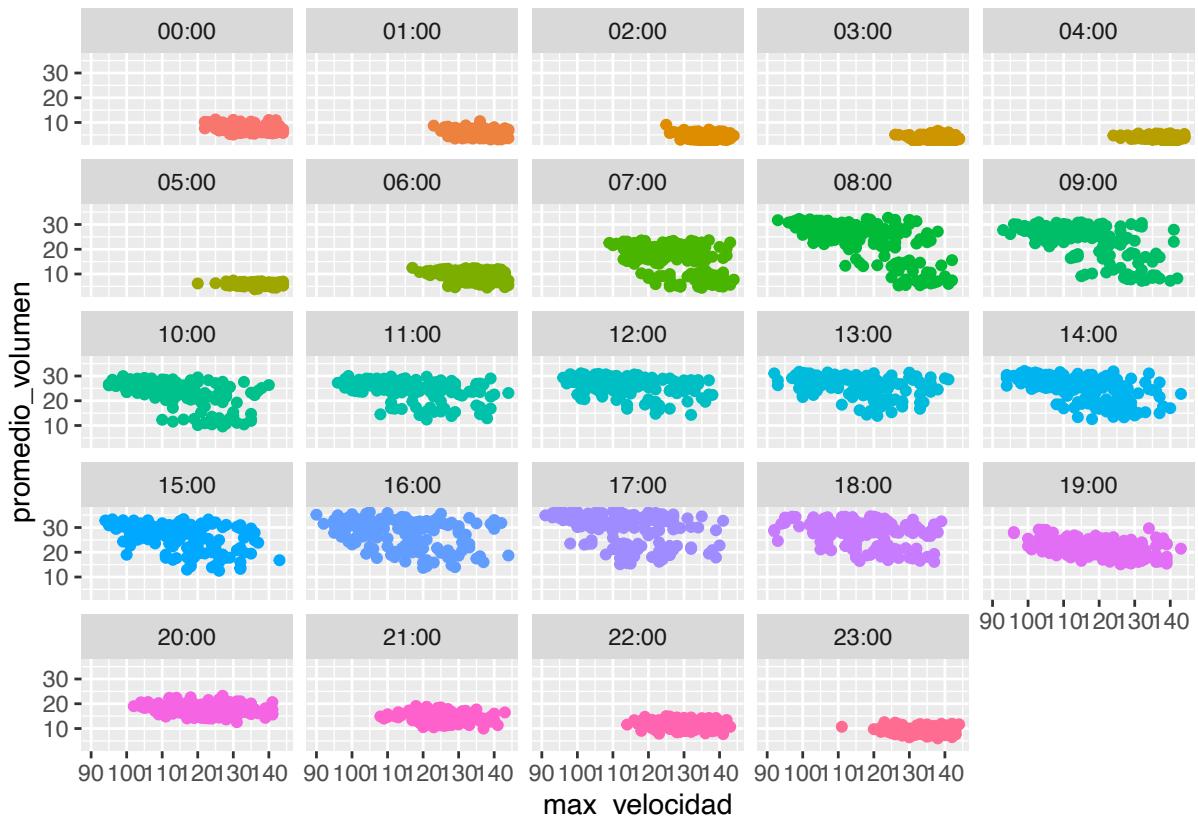
Variacion conjunta de velocidad y volumen



```
## <ggproto object: Class FacetWrap, Facet, gg>
##   compute_layout: function
##   draw_back: function
##   draw_front: function
##   draw_labels: function
##   draw_panels: function
##   finish_data: function
##   init_scales: function
##   map_data: function
##   params: list
##   setup_data: function
##   setup_params: function
##   shrink: TRUE
##   train_scales: function
##   vars: function
##   super:  <ggproto object: Class FacetWrap, Facet, gg>
```

Se observa una relación negativa, ya que cuanto mayor es el volumen, menor es el máximo de velocidad, y cuando el volumen es menor, los picos de velocidad tienden a ser mucho mayores.

Además, si separamos por hora del día podemos ver que los momentos de mayor volumen de tráfico suelen darse a la tarde, entre las 16:00 y las 18:00. También, los momentos donde el volumen es menor, y los picos de velocidad mayores suelen darse en la madrugada, entre las 2:00 y las 4:00 como muestra el siguiente gráfico.



Respecto a la “Variación conjunta de velocidad y volumen” creo que la motivación de hacerlo fue buena pero no supieron resolverla.

Resultados interesantes

Hemos visto en general un promedio de velocidad constante a lo largo de los con mayor volumen de tráfico durante los fines de semana casi debajo de los límites establecidos. No obstante existen registros de alta velocidad en masa con mayor frecuencia en la madrugada siendo registros que se toman en las zonas urbanas de Montevideo lo cual es llamativo

Hemos visto una ciudad con un volumen promedio de 17, si lo vemos barrio por barrio notaremos un volumen promedio entre 30 y 40 autos entre las 7 y las 18hs y muy poca circulación en la madrugada excepto el barrio Parque Rodo teniendo incluso mayor volumen de 0 a 4hs que el resto de las horas.

Hay asociación entre las dos variables, podrían haber realizado un modelo de regresión. Tal vez no se entiende lo que quieren decir

Modelo estadístico

Para el diseño del modelo, nos pareció interesante evaluar la interacción entre el volumen y la velocidad, además de otros factores planteados en las preguntas iniciales, como la hora o el día de la semana. Para esto, observamos que estas variables están claramente correlacionadas, por lo cual no es viable hacer un modelo de regresión ya que es necesaria la independencia de los errores. Por esto, concluimos en que es una mejor opción hacer un árbol de decisión, ya que nos permite observar esta dependencia con más claridad.

Predictir velocidad promedio de un sensor

Ya definido el tipo de modelo, nos resta definir la variable de respuesta y sus predictoras. Nos pareció que la mejor opción para ser variable de respuesta era la velocidad, ya que vemos que cada uno de los otros factores son condicionantes para esta variable.

Luego, nuestras variables predictoras serán el volumen, la hora y el día de la semana.

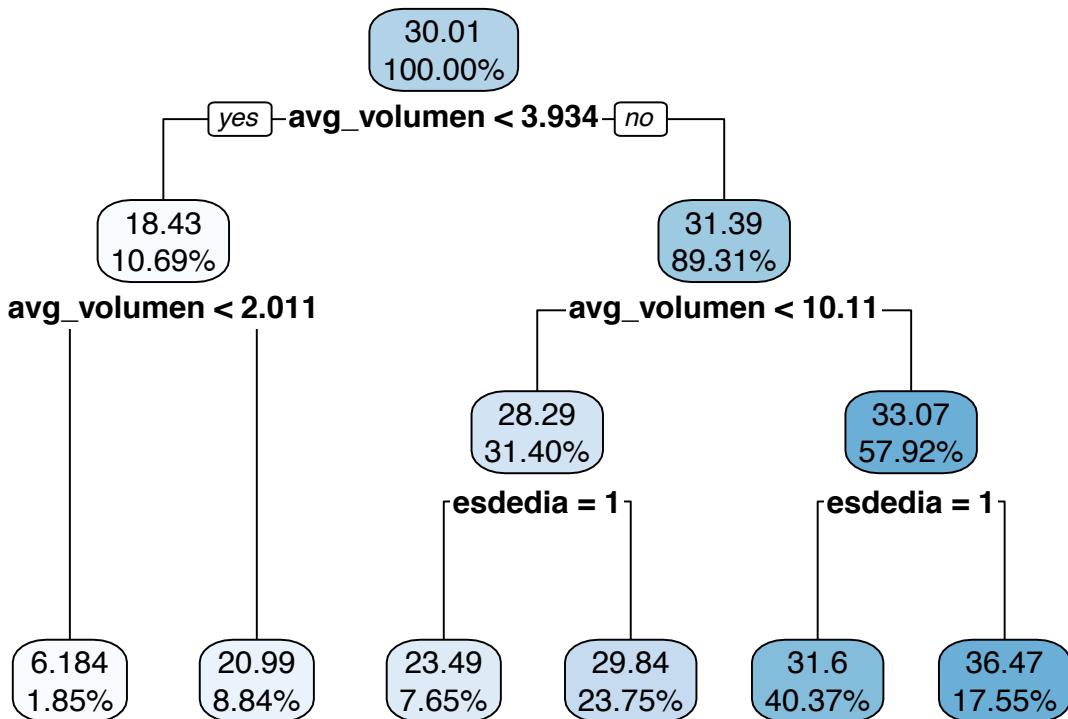
Para la hora y el día, decidimos seccionar las variables de forma binaria, ya que observamos en la velocidad una tendencia de comportamiento distinta en dos bloques bien definidos de cada variable. La hora estará seccionada en “día” y “noche” siendo “día” entre las 8:00 y las 20:00, y “noche” el caso contrario. Para el día de la semana, seccionaremos los datos en “fin de semana” y “día de semana”, ya que el tráfico suele comportarse de maneras diferentes en cada caso.

Para el volumen y la velocidad, tomamos el promedio para simplificar los datos. Los datos están agrupados por detector, hora, y día de la semana (Fin de semana o no).

```
## [1] 1083
```

Ya tenemos los datos, nos queda armar el arbol. Primero, tomaremos una muestra para crear los conjuntos de entrenamiento y prueba. La proporción será de un 70% para entrenamiento y un 30% para prueba. Antes de tomar la muestra fijaremos una semilla para poder analizar el mismo modelo de forma reproducible.

Al observar el modelo, notaremos que el volumen aparece repetidas veces seccionando los datos. Esto es porque, al ser continua, hay una variabilidad mucho más alta, y hay más casos para observar.



En concordancia con el comentario anterior, la variable volumen divide los datos en 4 categorías (Con límites 2.01; 3.93; 10.11) y la variable “esdedia” solo actúa una vez, mientras la del fin de semana ni siquiera es utilizada por el arbol.

Poedemos observar que los datos donde el volumen es menor a 3.93 es una minoría, ya que abarcan poco más de una décima parte. De todas formas, hay una división marcada en estos datos, ya que si el volumen es menor a 2, la velocidad suele ser de 6km/h, mientras que cuando el volumen es mayor a 2, la velocidad aumenta a casi 21km/h (Además es más significativa la cantidad de datos).

Respecto al modelo, hay salidas que no entendí muy bien que son, la interpretación del árbol esta correcta aunque mucha de las cosas se pueden ver en el gráfico. Cuando hablan de la “fiabilidad” del modelo podrían haber utilizado otras métricas, de todas maneras no es el error cuadrático medio es la raíz del error cuadrático medio, se ve en el código.

Cuando el volumen es mayor a 3.93 también hay un límite que demarca un comportamiento distinto entre los datos que superan y no este número, y es el 10.11. Los datos que tienen un volumen menor a 10.11, abarcan casi un tercio de los datos con una velocidad promedio de 28.29km/h, y los datos con volumen mayor a 10.11 abarcan casi un 58% de los datos con una velocidad promedio de 33 km/h. Se observa que la velocidad es mayor cuando el volumen es mayor a 10.11, aunque no es una diferencia tan significativa.

Una observación interesante es la diferencia de comportamiento entre los datos del día y de noche según el volumen. En todos los casos la velocidad es considerablemente mayor de noche que de día, pero la cantidad de datos observados es mayor de noche si el volumen es menor a 10, pero es mayor de día si el volumen es mayor a 10. Es decir, cuando el volumen es menor, hay más observaciones de noche, pero cuando hay mucho volumen de tráfico, hay menos observaciones de noche y de día hay muchas más.

Para saber la fiabilidad del modelo, calculamos el error cuadrado medio, y en base a eso fuimos ajustando los parámetros del modelo hasta llegar al actual, ya que es el que menor error tiene.

```
## [1] 9.881231
```

```
## [1] 11.08959
```

El modelo se aleja alrededor de esa cantidad de kilómetros de los datos reales. Hay un mayor error en el conjunto de prueba, ya que son menos valores.

Aplicación Shiny

Enlace *La hora podría haberse incluido con un selector, como está no queda claro si estas filtrando hasta la hora que esta seleccionada o la hora que se marca, en el mapa le podrían haber incluido un poco más de trabajo al tooltip e incluir que agrupe puntos dependiendo del zoom del mapa para ver realmente un mapa de calor. En la sección gráficos, hubiera estado bueno poder seleccionar más de un barrio y día de la semana, así como también más de una variable, Comentarios finales sería interesante ver la velocidad promedio y máxima al mismo tiempo*

En promedio Montevideo es una ciudad con velocidad promedio de casi 31 km/h, no podemos dejar pasar por alto la cantidad de excesos de velocidad a valores muy elevados que durante el fin de semana se producen mas. Con ese valor de velocidad promedio y mirando tambien el volumen promedio de 17, podemos llegar a decir que Montevideo es una ciudad bastante lenta teniendo en consideración que las mediciones fueron hechas incluso en avenidas donde mas se circula, mucho mas en las famosas “hora pico” entre las 16 y 18hs

Referencias

- Detomasi, Richard. 2023. “geouy: Geographic Information of Uruguay.” <https://github.com/RichDeto/geouy>.
- Hvitfeldt, Emil. 2021. *Paletteer: Comprehensive Collection of Color Palettes*. <https://github.com/EmilHvitfeldt/paletteer>.
- Liaw, Andy, and Matthew Wiener. 2002. “Classification and Regression by randomForest.” *R News* 2 (3): 18–22. <https://CRAN.R-project.org/doc/Rnews/>.
- Milborrow, Stephen. 2022. *Rpart.plot: Plot 'Rpart' Models: An Enhanced Version of 'Plot.rpart'*. <https://CRAN.R-project.org/package=rpart.plot>.
- Pebesma, Edzer, and Roger Bivand. 2023a. *Spatial Data Science: With applications in R*. Chapman and Hall/CRC. <https://r-spatial.org/book/>.
- Pebesma, Edzer, and Roger S. Bivand. 2023b. *Spatial Data Science with Applications in R*. Chapman & Hall. <https://r-spatial.org/book/>.
- R Special Interest Group on Databases (R-SIG-DB), Hadley Wickham, and Kirill Müller. 2022. *DBI: R Database Interface*. <https://CRAN.R-project.org/package=DBI>.

- Therneau, Terry, and Beth Atkinson. 2022. *Rpart: Recursive Partitioning and Regression Trees*. <https://CRAN.R-project.org/package=rpart>.
- Wickham, Hadley. 2021. *Mastering Shiny: Build Interactive Apps, Reports, and Dashboards Powered by r*. O'Reilly.
- . 2023. *Modelr: Modelling Functions That Work with the Pipe*. <https://CRAN.R-project.org/package=modelr>.
- WICKHAM, HADLEY. 2023. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'REILLY MEDIA.
- Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. <https://doi.org/10.21105/joss.01686>.
- Wickham, Hadley, Jeroen Ooms, and Kirill Müller. 2023. *RPostgres: Rcpp Interface to PostgreSQL*. <https://CRAN.R-project.org/package=RPostgres>.

ESTE PROYECTO TENÍA VARIOS DESAFIOS Y HAN PODIDO SUPERAR MUCHOS DE ELLOS SIN EMBARGO EL TRABAJO ESCRITO QUEDÓ BASTANTE DESORDENADO Y LA DEFENSA TAMBIÉN. HAY COSAS INTERESANTES QUE LOGRARON COMO EL ANÁLISIS DE VELOCIDAD MÁXIMA Y LA APERTURA POR HORA Y DÍAS Y CREO QUE CON MAS TIEMPO HUBIERAN REALIZADO UN TRABAJO MUCHO MEJOR. CREO QUE APRENDIERON BASTANTE EN EL TRABAJO Y HAY QUE SEGIR ESTUDIANDO ESTADÍSTICA PARA MEJORAR! PUNTOS 75/100