

Datos de Transito

Ciencia de datos con R

Iván Arriola, Federico Miquelerena, Damián Rovetta

12-07-2023

Introducción

Los datos utilizados en este proyecto fueron sacados de Catalogo de Datos Abiertos de **gub.uy**. Los mismos corresponden a los datos de transito de la ciudad de Montevideo en el periodo de Enero 2021 a Mayo 2023. 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.

Base de datos para el proyecto

Para el proyecto se utilizo una base de datos que contiene los tres dataset mencionados anteriormente.

Se combinaron los datos del dataset dando origen a 3 tablas: ### Tabla: fct_registro

- ▶ Cantidad de datos: 85386695.
- ▶ Variables de la tabla:
 - ▶ *id_registro*: 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 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

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

Base de datos para el proyecto

Y una tercer tabla que contiene los datos de las fechas de los registros.

Tabla: d_date

Algunas de las variables de esta tabla son

- ▶ 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
 - ▶ *day_of_week*: Numérico. Dia 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. Dia del cuatrimestre.
 - ▶ *day_of_year*: Numérico. Dia del año, del 1 al 366.
 - ▶ *month_actual*: Numérico. Mes del año tomado como numero, enero como 1, febrero como 2 y así sucesivamente.
 - ▶ *year_actual*: Numérico. Indica el año.
 - ▶ *mmyyyy*: Numérico. Secuencia de caracteres que indica el mes y el año en formato MMYYYY
 - ▶ *mmddyyyy*: 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 dia de la semana sabado o domingo, FALSE en caso contrario.
 - ▶ *feriado*: Lógico. TRUE si la fecha correspondiente coincide con dias feriados en Uruguay, FALSE en caso contrario.

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.

Datos Faltantes o Nulos

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.

Datos Faltantes o Nulos

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
```

Datos Faltantes o Nulos

```
##   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
```

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.

Ubicación de los sensores

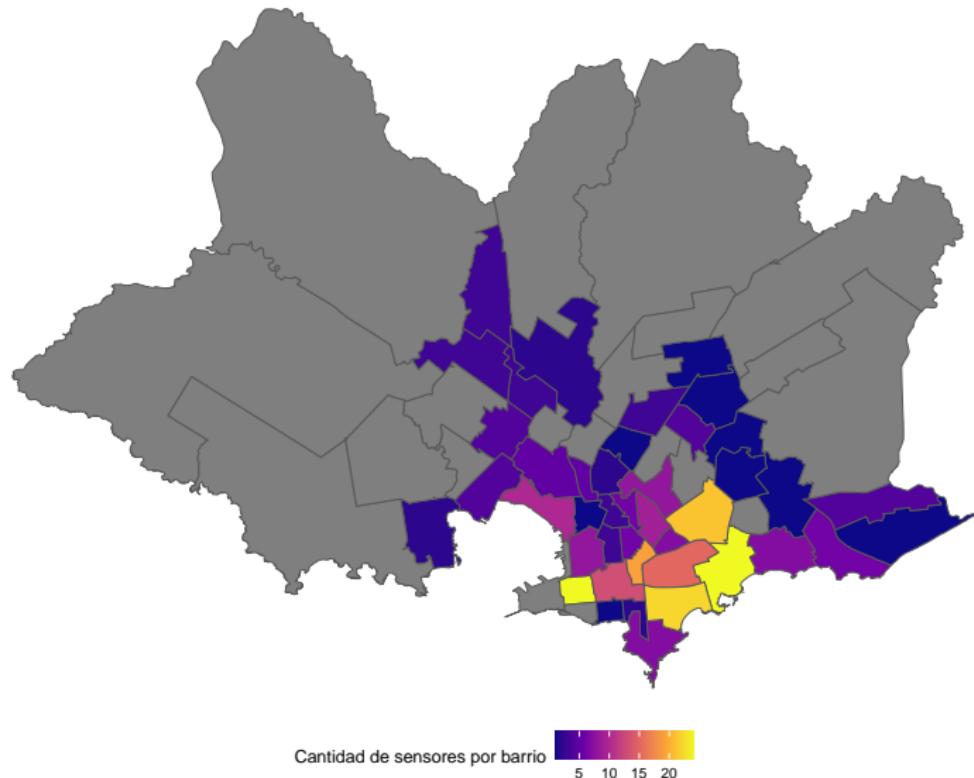
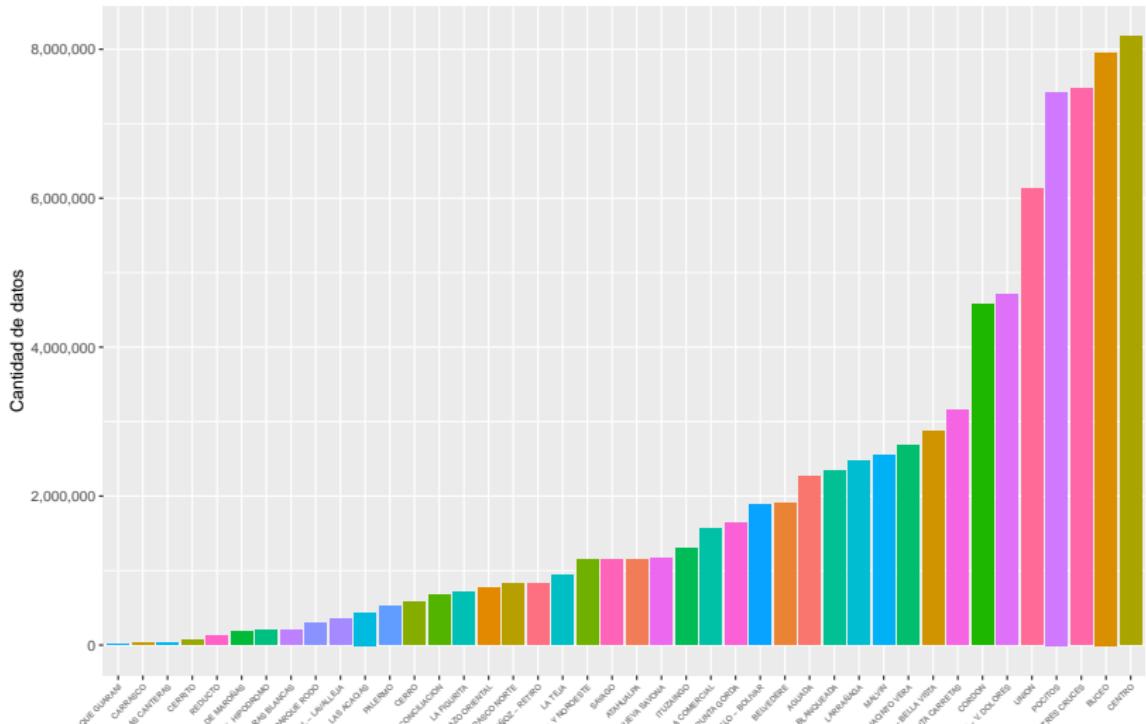


Figura 1: Mapa de Montevideo con cantidad de sensores por barrio.

Ubicacion de los sensores

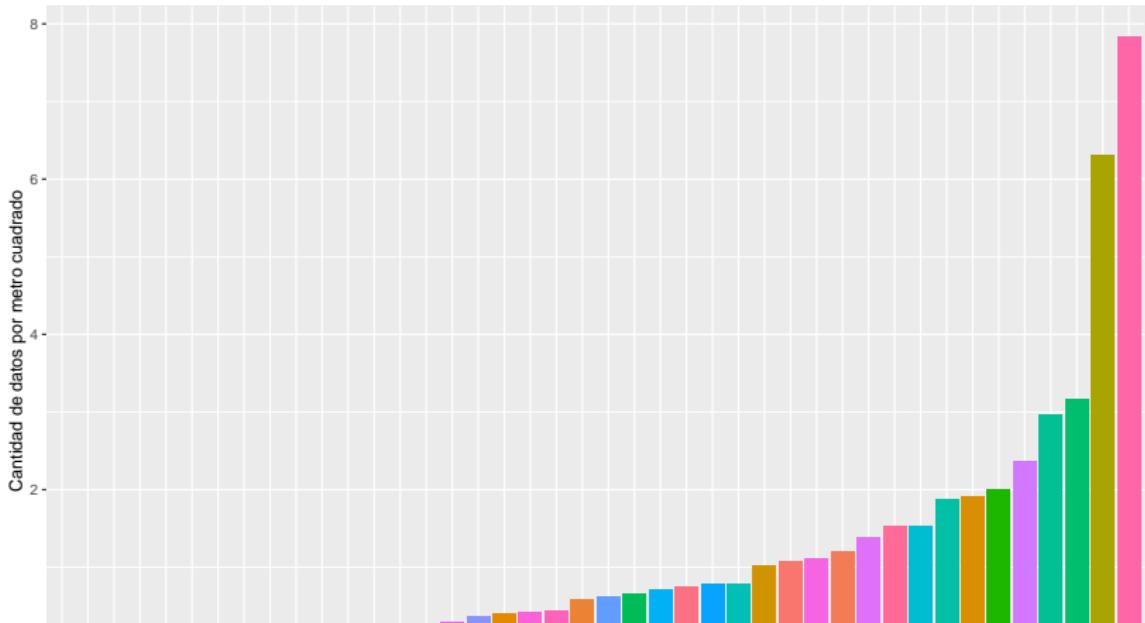
Ahora quiero mostrar la cantidad de datos que tenemos por ubicacion, para ver si hay alguna ubicacion en particular que tenga mas o menos datos que las demás.



Ubicacion de los sensores

Ahora me interesaria 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.



Principales variables

Las variables que se van a analizar son las siguientes: - *volume*: *Numérico*. Cantidad de vehículos que pasaron en los últimos 5 minutos. - *volumen_hora*: *Numérico*. Cantidad de vehículos que pasaron en la ultima hora. - *velocidad*: *Numérico*. Velocidad promedio de los vehículos registrados en los ultimos 5 minutos. Unidad en km/h

Velocidad

Resumen de la variable velocidad

```
##      cant_0 minimo primer_cuartil mediana tercer_cuartil maximo promed
## 1 8873659      0          22       32          41     144 31.312
```

Velocidad

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

Velocidad

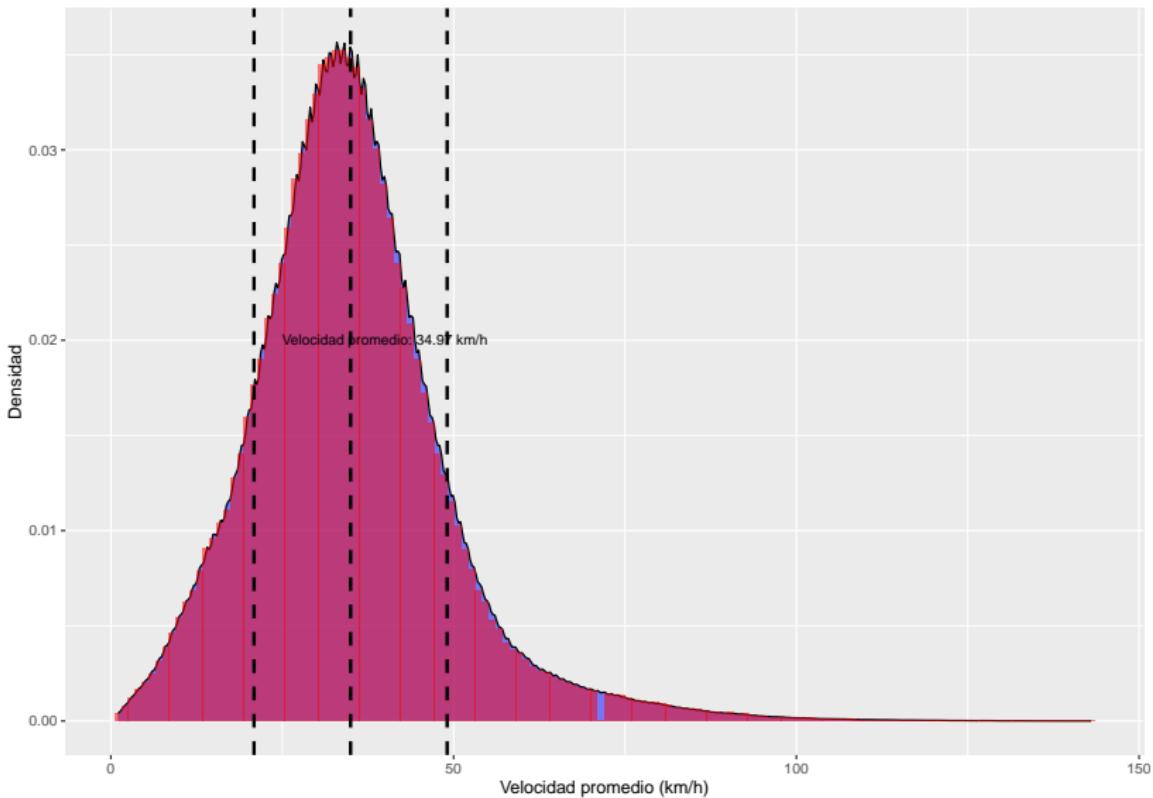


Figura 4: Distribucion de velocidad promedio

Volumen

Resumen de la variable volumen

minimo, maximo, promedio, desvio, cuartiles

```
##      cant_0 minimo cuartil_1 mediana cuartil_3 maximo promedio    desvi
## 1 8873659      0        3       11       26      659 17.27588 19.1834
```

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.

Volumen

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

Volumen

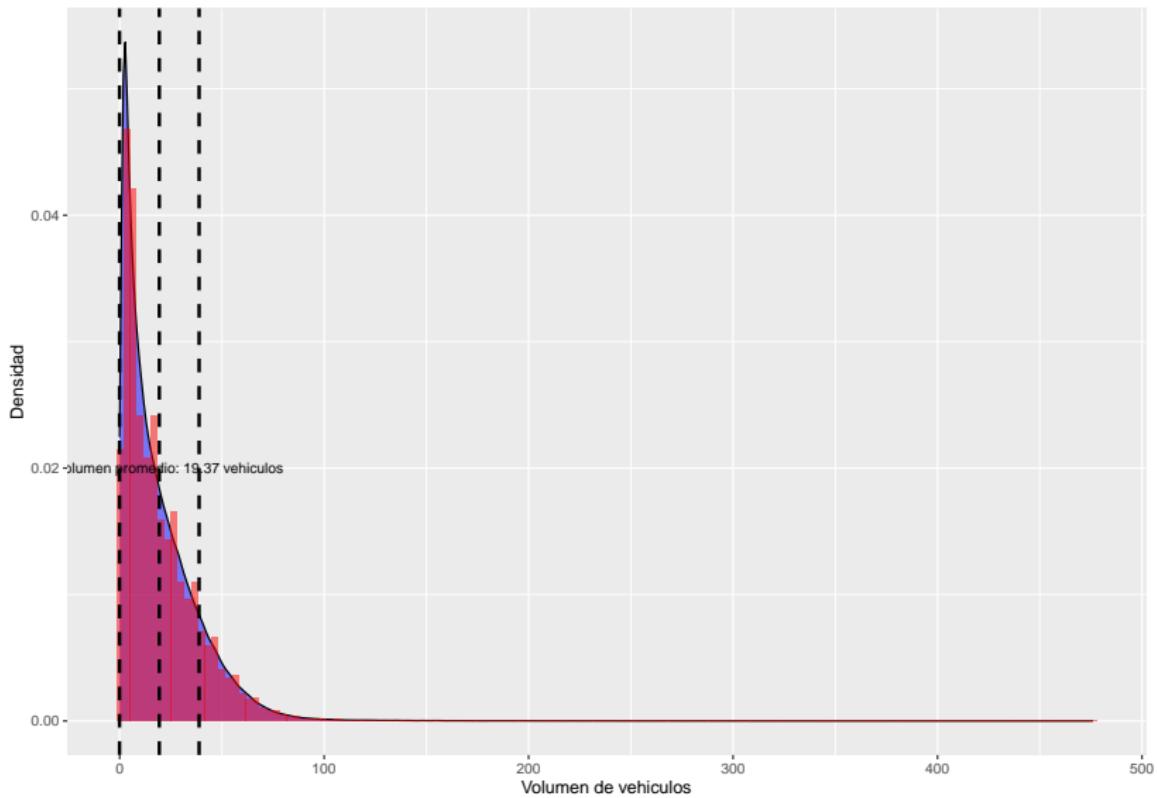


Figura 5: Distribucion de volumen

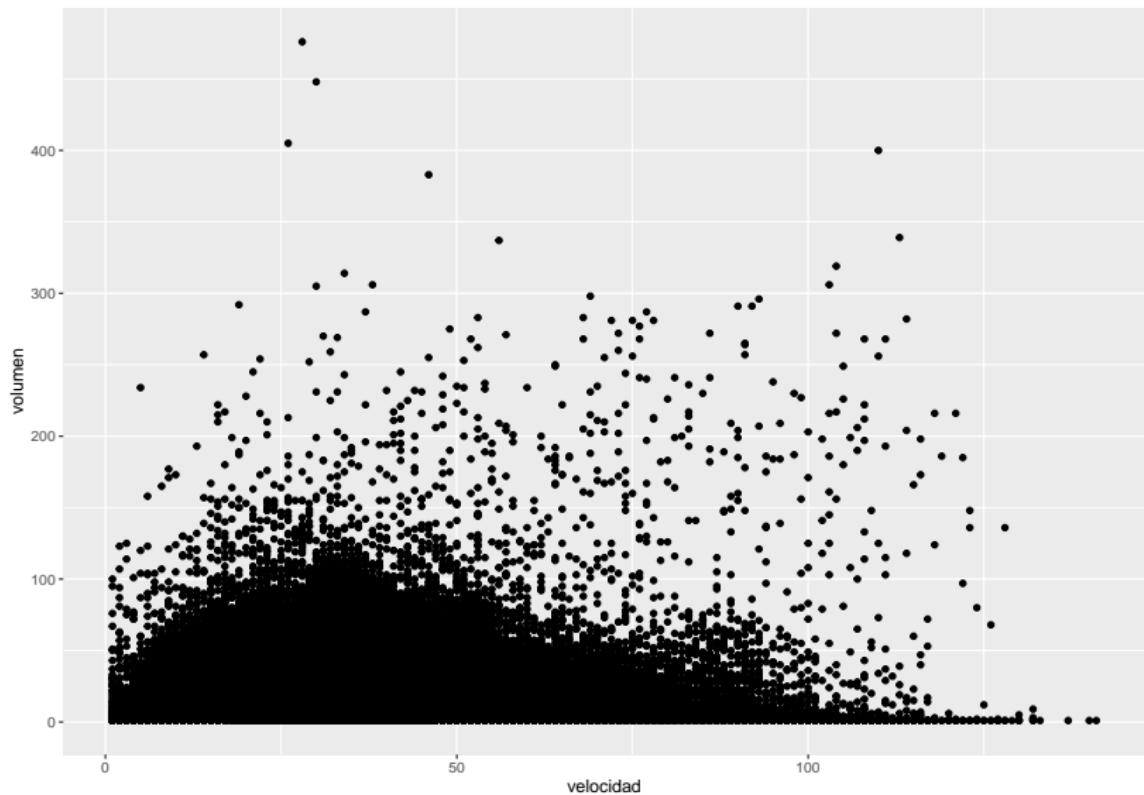
Preguntas de Investigación

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 cometan 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



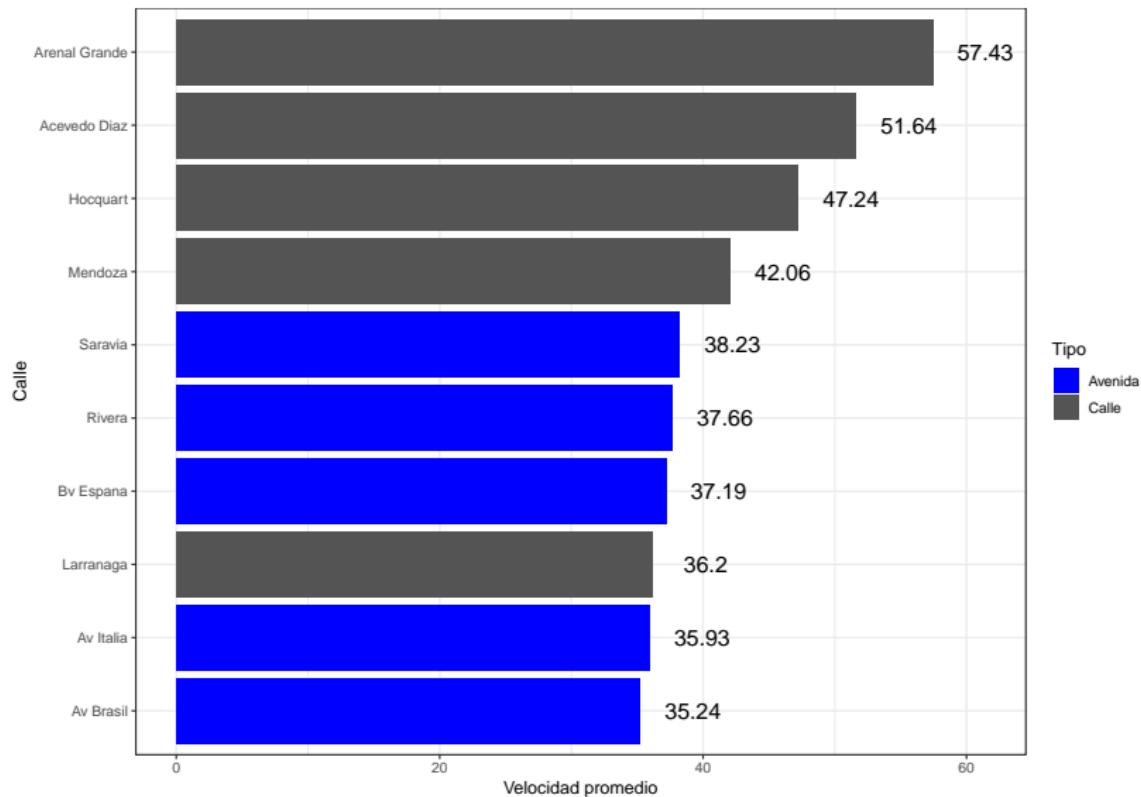
1. ¿Existe alguna correlación entre el volumen y la velocidad?

```
##           velocidad      volumen
## velocidad  1.00000000 -0.03175882
## volumen    -0.03175882  1.00000000
```

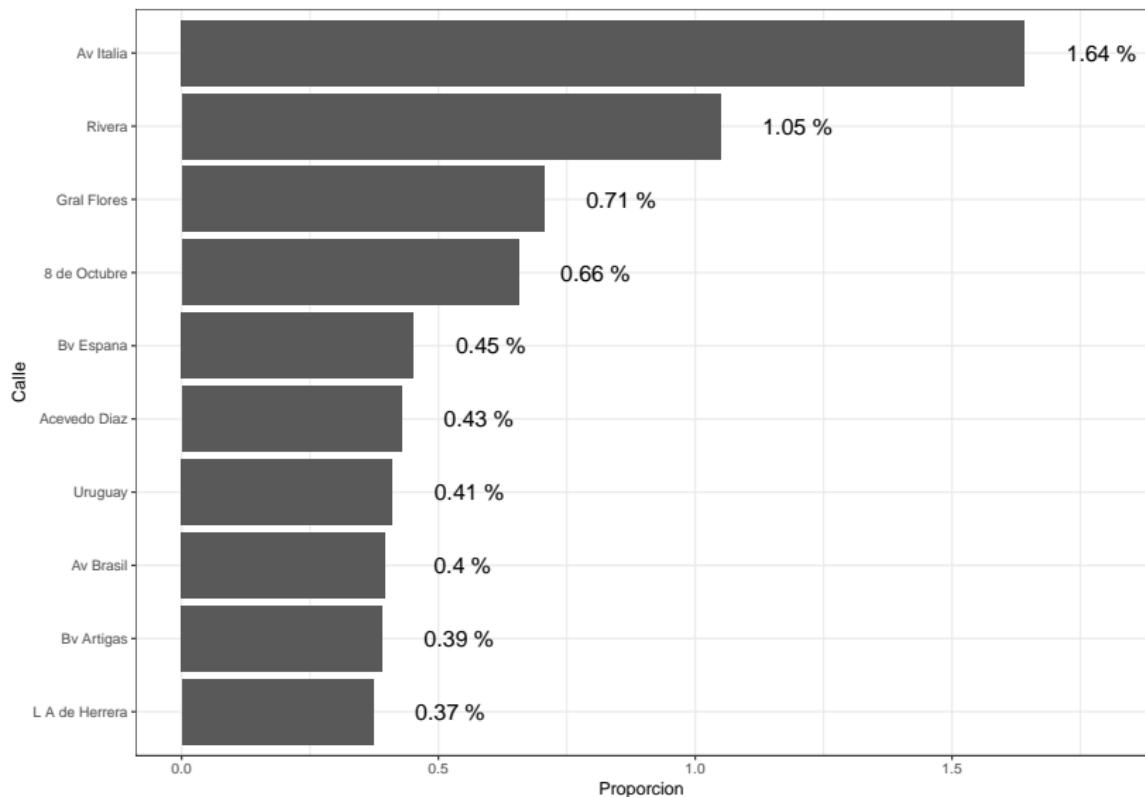
La correlacion dio -0.03, lo que indica que no hay una correlacion lineal entre las variables.

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

Pasemos a investigar las calles con mayor promedio de velocidad.



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



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

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.

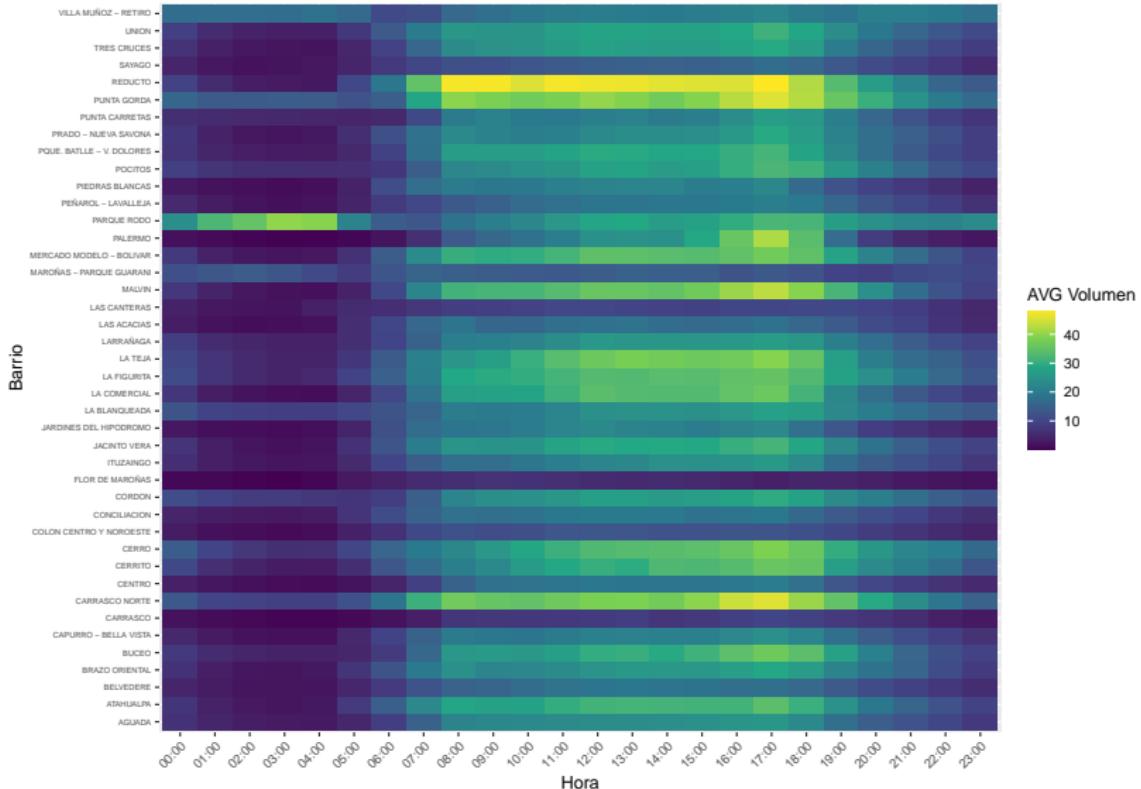
3. ¿Cómo va variando el volumen y velocidad medidos a traves de el tiempo?

Volumen

Primero observemos como va variado volumen y volumen hora

Volumen

Variacion por hora del dia

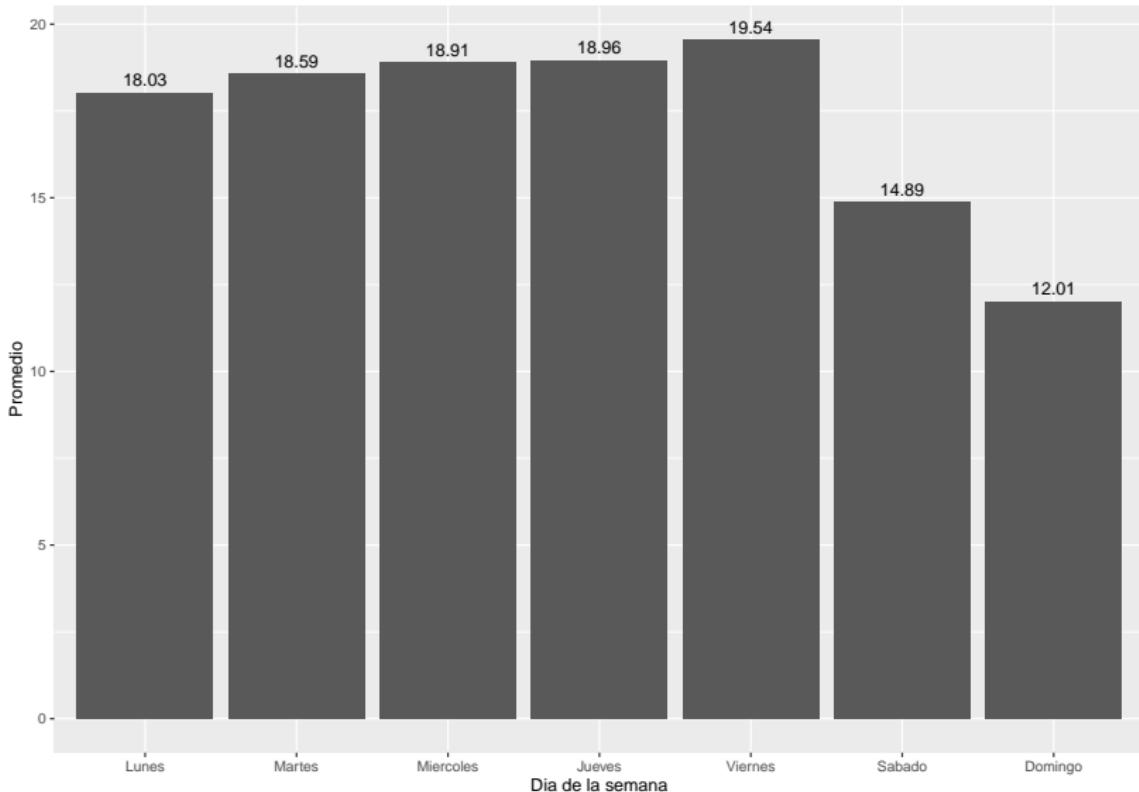


Volumen

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.

Volumen



Volumen

Se puede notar un comportamiento casi idéntico en la cantidad de

Volumen

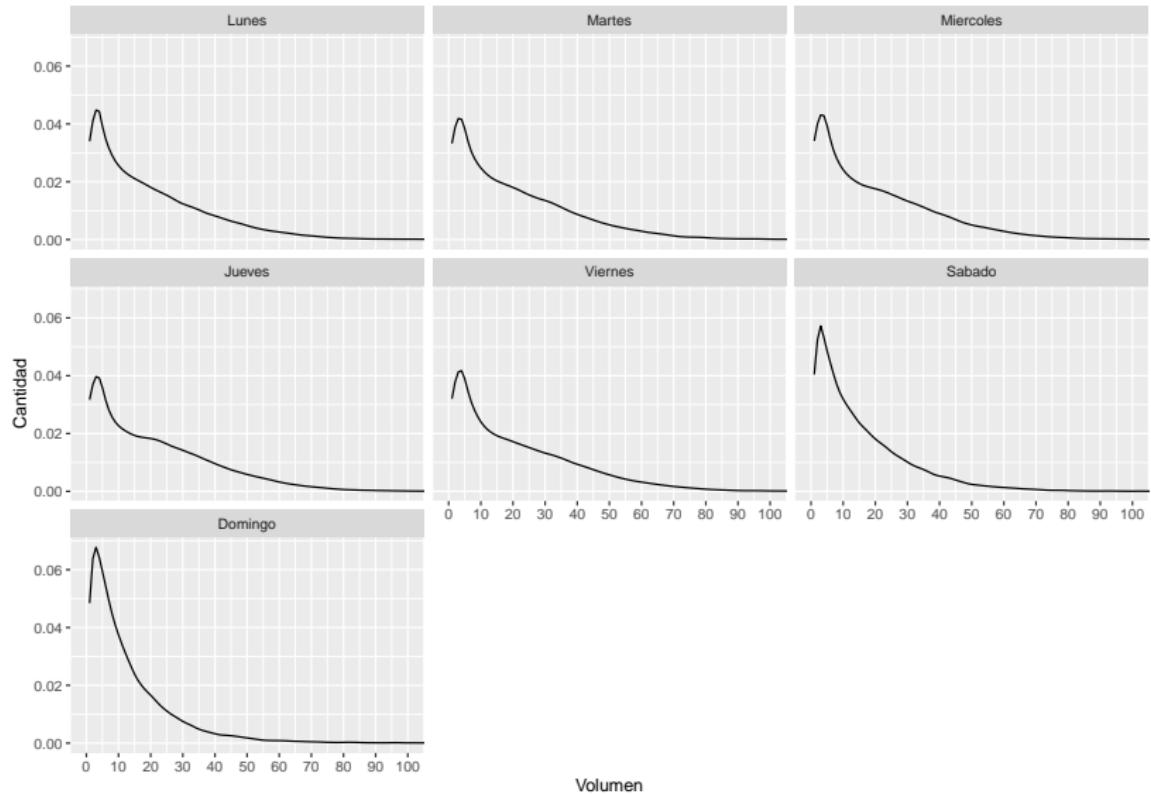


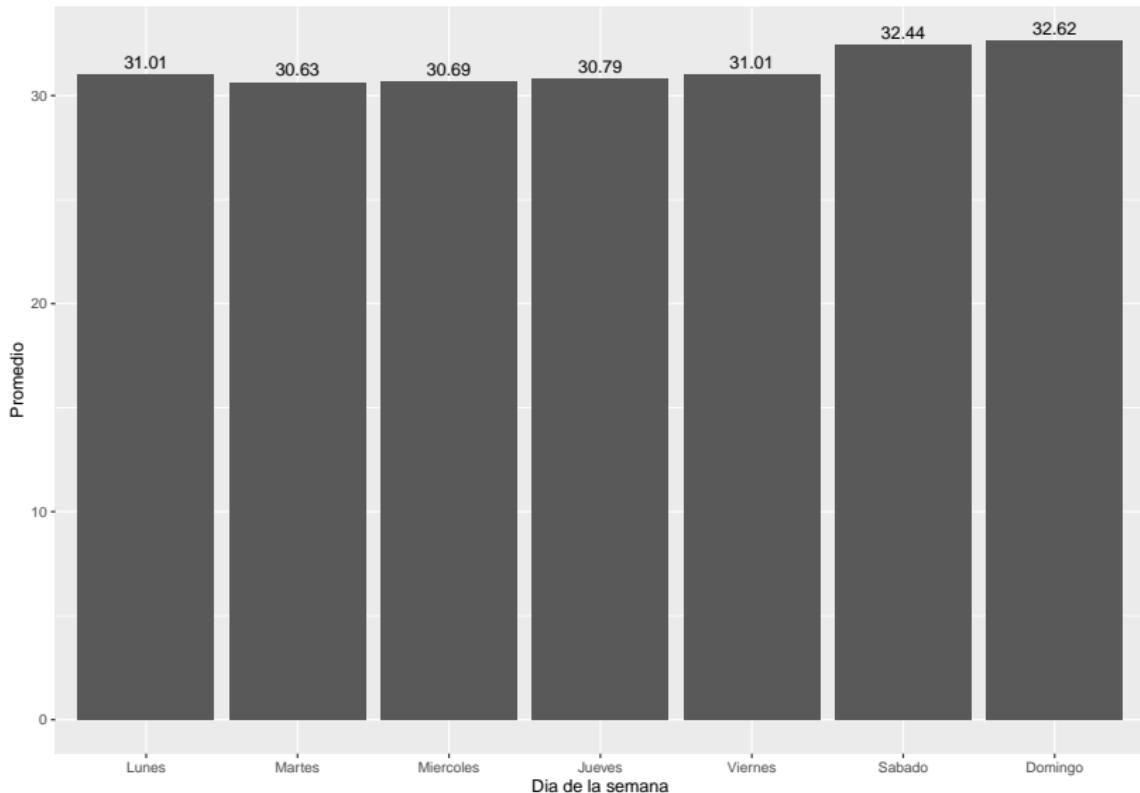
Figura 8: Densidad de Volumen

Velocidad

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.

Velocidad



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.

Velocidad

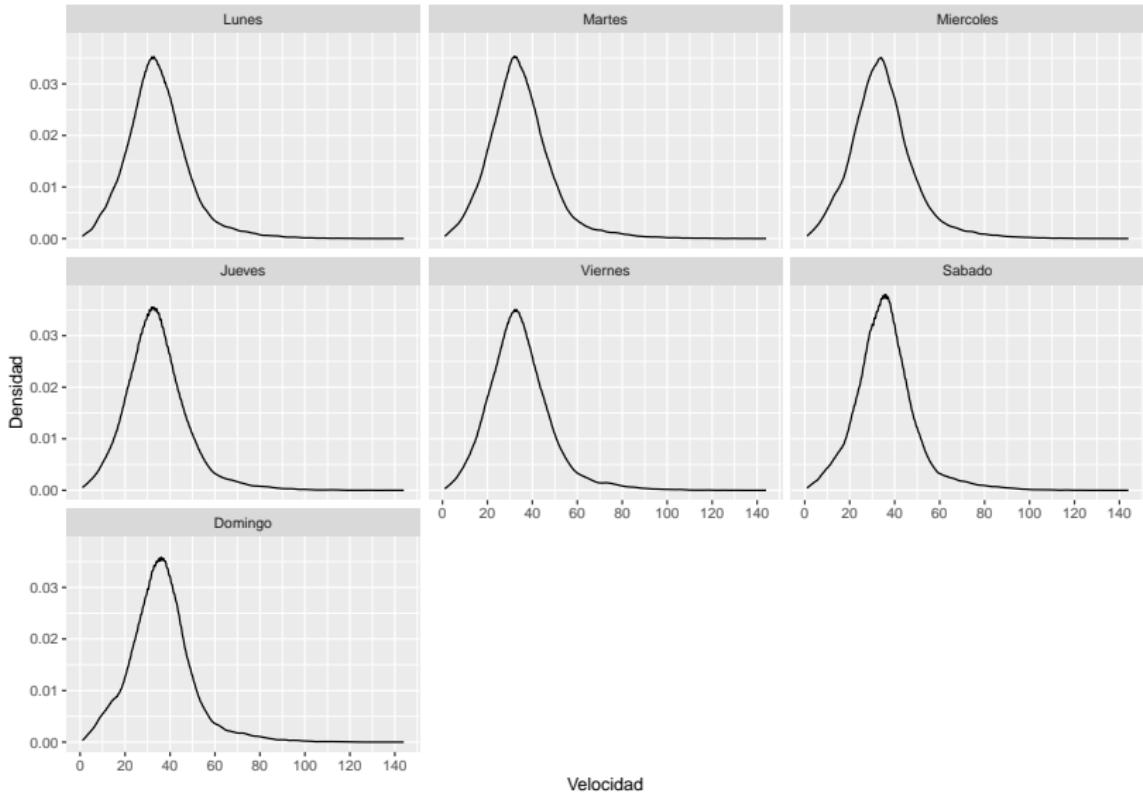


Figura 9: Densidad de Velocidad

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.

Velocidad

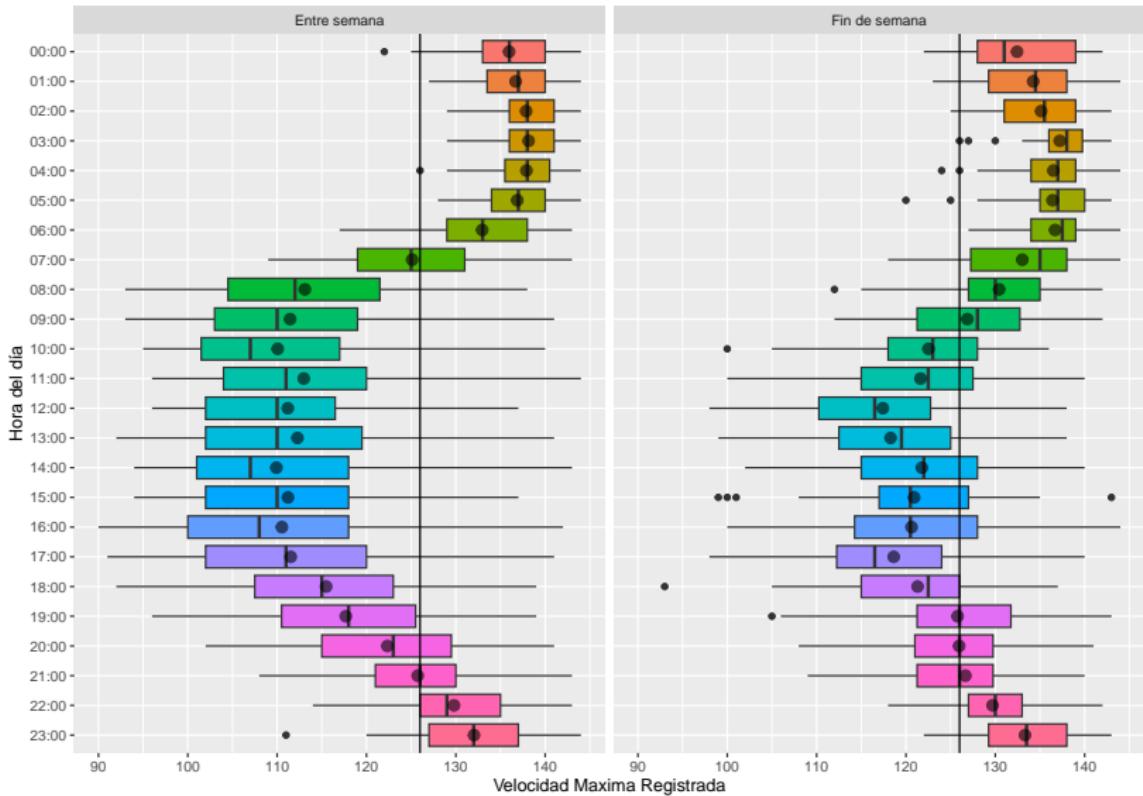


Figura 10: Distribucion maxima velocidad por hora

Velocidad

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.

Velocidad

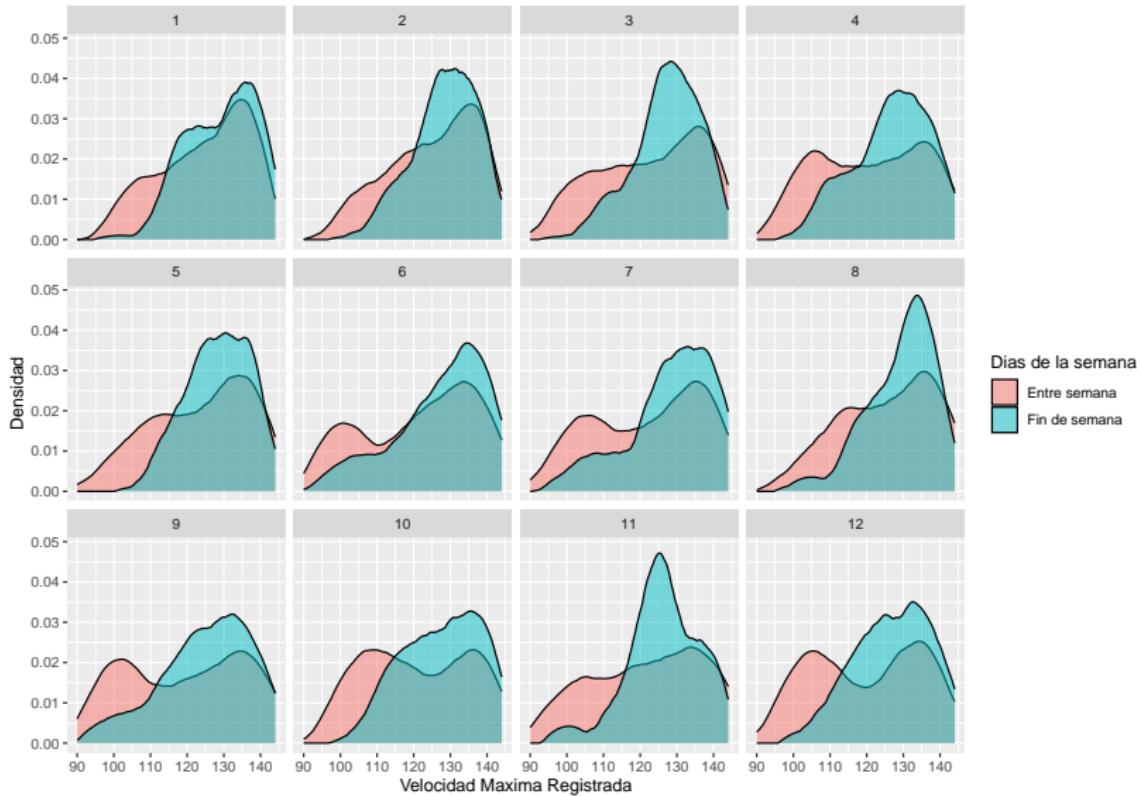


Figura 11: Densidad de la velocidad por meses

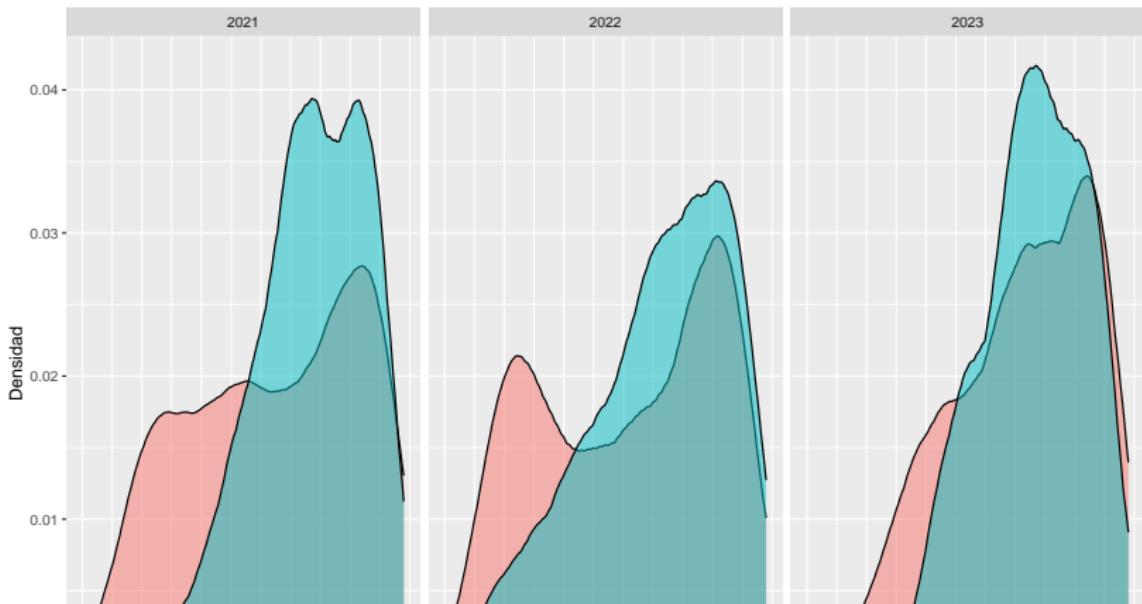
Velocidad

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.



Velocidad

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.

Velocidad

Variacion conjunta de velocidad y volumen



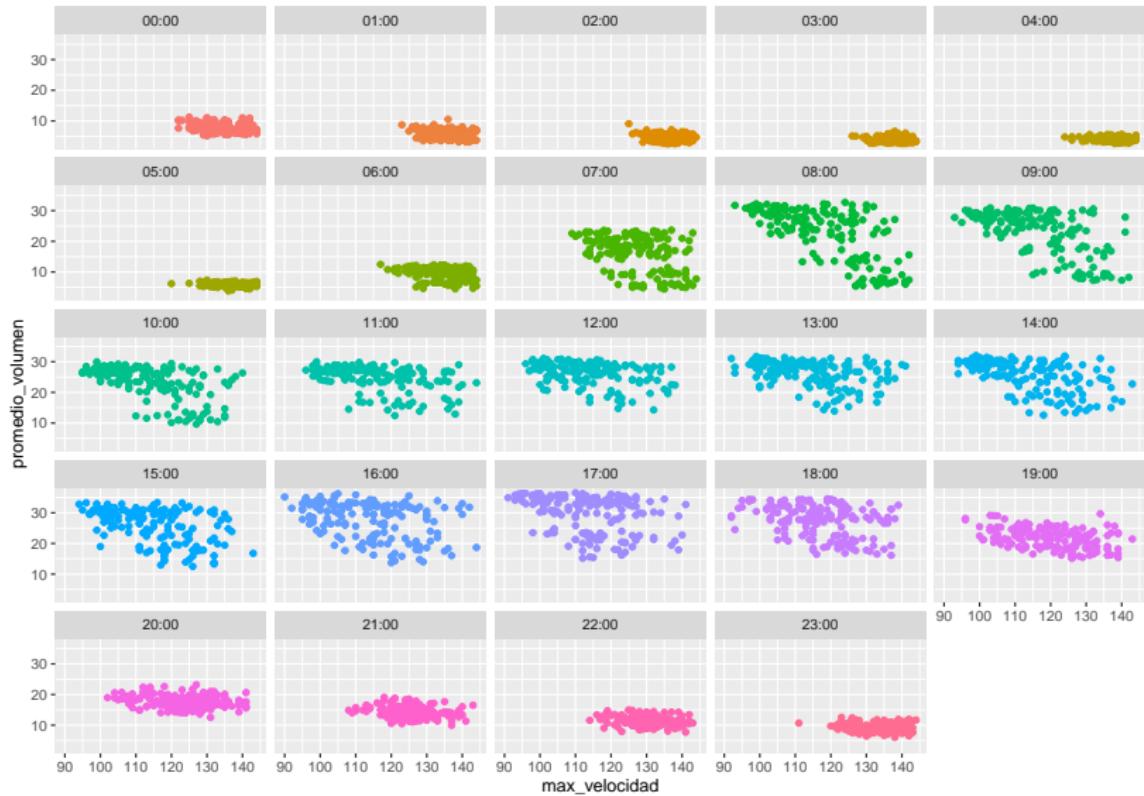
```
## <ggproto object: Class FacetWrap, Facet, gg>
```

Velocidad

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.

Velocidad



Resultados interesantes

Hemos visto en general un promedio de velocidad constante a lo largo de los con mayor volumen de trafico durante los fines de semana casi debajo de los limites 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.

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 arbol de decisión, ya que nos permite observar esta dependencia con más claridad.

Predicir 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.

Predicir velocidad promedio de un sensor

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).

Predicir velocidad promedio de un sensor

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

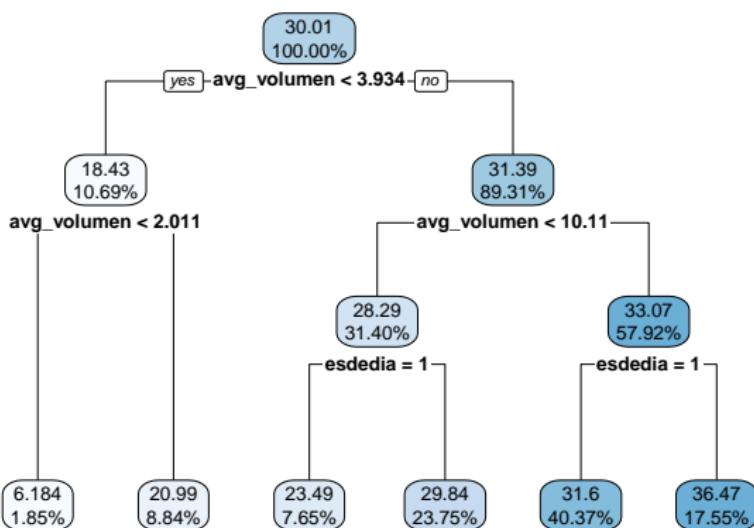
Predicir velocidad promedio de un sensor

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.

Predicir velocidad promedio de un sensor

Predicir velocidad promedio de un sensor

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.



Predicir velocidad promedio de un sensor

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 “esdia” 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).

Predicir velocidad promedio de un sensor

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.

Predicir velocidad promedio de un sensor

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.

Predicir velocidad promedio de un sensor

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

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

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

En la aplicacion hay tres pestañas:

- ▶ Mapa
- ▶ Graficos

Mapa

En esta sección se presenta un mapa de Montevideo, obtenido del paquete geouy. Para su realización se utilizó el paquete leaflet. En este mapa se puede visualizar ciertos resúmenes de los datos de velocidad y cantidad de vehículos.

Datos de Trafico

Variables a mostrar

 Barrio
 Sensor

Seleccione una hora



0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22

Seleccione una variable

 Volumen Promedio
 Velocidad Promedio
 Velocidad Máxima
 Volumen Máximo

Seleccione un dia de la semana

 Lunes
 Martes
 Miércoles
 Jueves
 Viernes
 Sábado
 Domingo

Figura 12: Mapa

El mapa tiene dos capas posibles de visualizar, una es por barrios y la otra por sensores.

En cada una de estas capas puedes filtrar los datos por día de la semana y por hora. También puedes ir cambiando la variable de resumen que se muestra en el mapa.

Grafico

En esta sección se presenta un gráfico de barras que muestra la variable de resumen que se seleccione por hora. También se filtra por día de la semana y por barrio.

