Accidentes de tránsito en New York

jajaimes4@poligran.edu.co

# 1. Introduccion

La ciudad de New York es una de las mas habitadas del mundo con mas de 8M de habitantes hacias el 2022, como se menciona en DataUsa.io, la diversidad de su poblacion varia con un 85% de ellos serian ciudadaos nacionales y en su mayoria de ellos de identidada blanca, vale la pena mencionar que un 29% de ellos son hispano parlantes y ademas de ello mas conocida como la ciudad que nunca duerme.

En terminos economicos, mas de 4.11M de personas hacen parte de la fuerza laboral para el 2022. Llama la atencion este indicador porque ano tras ano esta en decadencia con un ritmo del declive del -1.34%.

los accidentes de trafico representan un problema de gran importancia en diferentes niveles, con repercusiones tanto economicas como sociales segun Lucas et al. (2024) . En espana por ejemplo represetan la mayor causa de mortalidad externa.

La pregunta que se plantea: se pueden establecer tendencias de crecimientos de los accidentes de transito en la ciudad de New York, que conlleve a definir politicas publicas y con ellos evitar que ciertos indicadores economicos no sean afectados?

# 2. Metodologia

El conjunnto de datos a analisar se llama a accidentes, cuenta con mas observaciones y tiene columnas tanto cualitativas como cuantitativas. Como parte del analisis exploratoria se intentaran contestar las siguientes preguntas: ¿Se ha incrementado el número de accidentes en el periodo de observación? ¿Qué podemos concluir acerca del número de accidentes por mes? ¿Existen patrones horarios en los accidentes? ¿Existen patrones en los accidentes según el día de la semana? ¿Existen patrones accidentales por vecindario? ¿Existen patrones horarios en la accidentalidad por vecindario? ¿Cuáles son las causas más comunes de los accidentes? ¿Cuáles son los tipos de vehículos involucrados en más accidentes? ¿Pueden existir patrones, de factores o vehículos, diferencial para vecindarios?. Para darle respuesta a las preguntas anteriores necesario la manipulaciom de los datos con R y la estructura de ello, a continuacion +

install.packages('flextable')

also installing the dependencies ‘fontBitstreamVera’, ‘fontLiberation’, ‘gfonts’, ‘fontquiver’, ‘zip’, ‘gdtools’, ‘officer’  
  
  
Updating HTML index of packages in '.Library'  
  
Making 'packages.html' ...  
 done

library("ggplot2")  
library("dplyr")  
library('tidyr')  
library('reshape2')  
library('flextable')  
  
  
setwd("./")

Attaching package: ‘dplyr’  
  
  
The following objects are masked from ‘package:stats’:  
  
 filter, lag  
  
  
The following objects are masked from ‘package:base’:  
  
 intersect, setdiff, setequal, union  
  
  
  
Attaching package: ‘reshape2’  
  
  
The following object is masked from ‘package:tidyr’:  
  
 smiths

d <- read.csv2("./datasets/accidents.csv", sep = ";", header = TRUE)

d$DATE <- as.Date.character(d$DATE, "%m/%d/%Y")  
d$M.Y <- strftime(d$DATE, "%y-%m")  
d$TIME <- strptime(d$TIME, "%H:%M")  
d$T <- strftime(d$TIME, "%H")  
d$D <- weekdays(d$DATE)  
d$D <- factor(d$D)  
  
head(d)

A data.frame: 6 × 27

|  | DATE <date> | TIME <dttm> | BOROUGH <chr> | ZIP.CODE <chr> | LATITUDE <chr> | LONGITUDE <chr> | ON.STREET.NAME <chr> | NUMBER.OF.PEDESTRIANS.INJURED <int> | NUMBER.OF.PEDESTRIANS.KILLED <int> | NUMBER.OF.CYCLIST.INJURED <int> | ⋯ ⋯ | CONTRIBUTING.FACTOR.VEHICLE.5 <chr> | COLLISION\_ID <int> | VEHICLE.TYPE.CODE.1 <chr> | VEHICLE.TYPE.CODE.2 <chr> | VEHICLE.TYPE.CODE.3 <chr> | VEHICLE.TYPE.CODE.4 <chr> | VEHICLE.TYPE.CODE.5 <chr> | M.Y <chr> | T <chr> | D <fct> |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | NA | NA | BRONX | 10454.0 | 40.808987 | -73.911316 |  | 0 | 0 | 0 | ⋯ |  | 3988123 | Sedan |  |  |  |  | NA | NA | NA |
| 2 | NA | NA | BROOKLYN | 11236.0 | 40.636005 | -73.91251 | FLATLANDS AVENUE | 1 | 0 | 0 | ⋯ |  | 3987962 | Sedan |  |  |  |  | NA | NA | NA |
| 3 | NA | NA | QUEENS | 11101.0 | 40.75549 | -73.93953 |  | 0 | 0 | 0 | ⋯ |  | 4193132 | Sedan |  |  |  |  | NA | NA | NA |
| 4 | NA | NA | QUEENS | 11367.0 |  |  | MAIN STREET | 0 | 0 | 1 | ⋯ |  | 3985962 | Bike | Station Wagon/Sport Utility Vehicle |  |  |  | NA | NA | NA |
| 5 | NA | NA | BRONX | 10468.0 | 40.868336 | -73.90127 |  | 0 | 0 | 0 | ⋯ |  | 4192111 | Sedan | Sedan |  |  |  | NA | NA | NA |
| 6 | NA | NA | QUEENS | 11354.0 |  |  |  | 0 | 0 | 0 | ⋯ |  | 4191802 | Sedan |  |  |  |  | NA | NA | NA |

d

dt <- d %>% select(M.Y, T, D, BOROUGH)  
  
dt <- data.frame(  
 Columns = names(dt),  
 Type = sapply(dt, class),  
 Observations = sapply(dt, length)  
)  
  
ft <- flextable(dt)  
ft <- theme\_vanilla(ft)  
ft

a flextable object.  
col\_keys: `Columns`, `Type`, `Observations`   
header has 1 row(s)   
body has 4 row(s)   
original dataset sample:   
 Columns Type Observations  
M.Y M.Y character 238522  
T T character 238522  
D D factor 238522  
BOROUGH BOROUGH character 238522

## 2.1 ¿Se ha incrementado el número de accidentes en el periodo de observación?

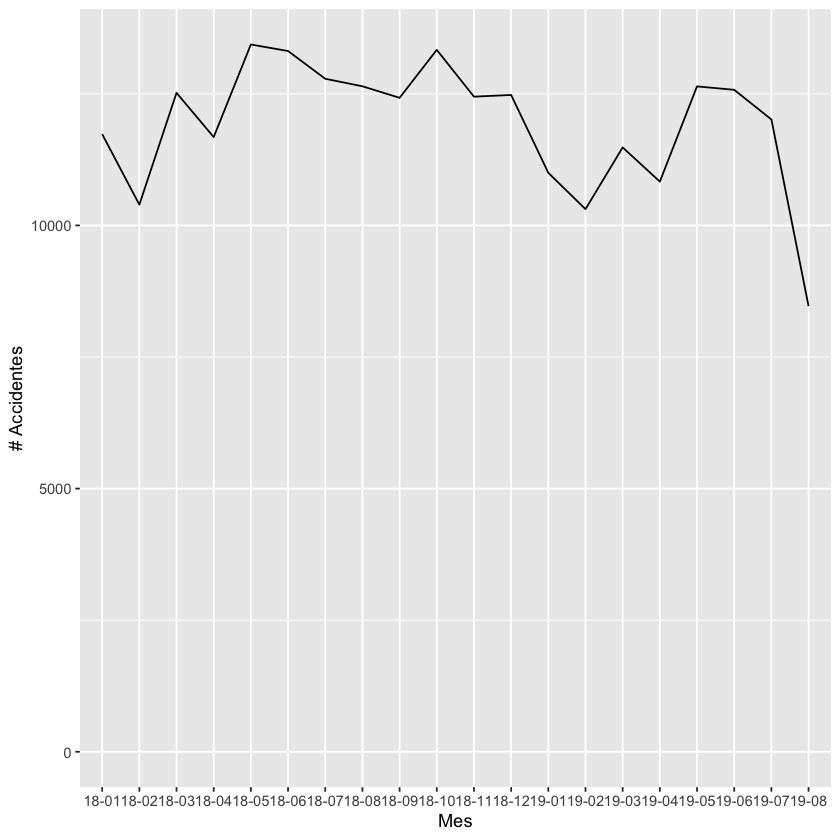
De acuerdo con el grafico, la tendendia es a la baja en los ultimos dos ano lo que signigica que esta desendiendo. Con una media de y una desviacion estandar de

d %>% count(M.Y) %>% summarise(m = mean(n), ds = sd(n))

A data.frame: 1 × 2

| m <dbl> | ds <dbl> |
| --- | --- |
| 11926.1 | 1232.317 |

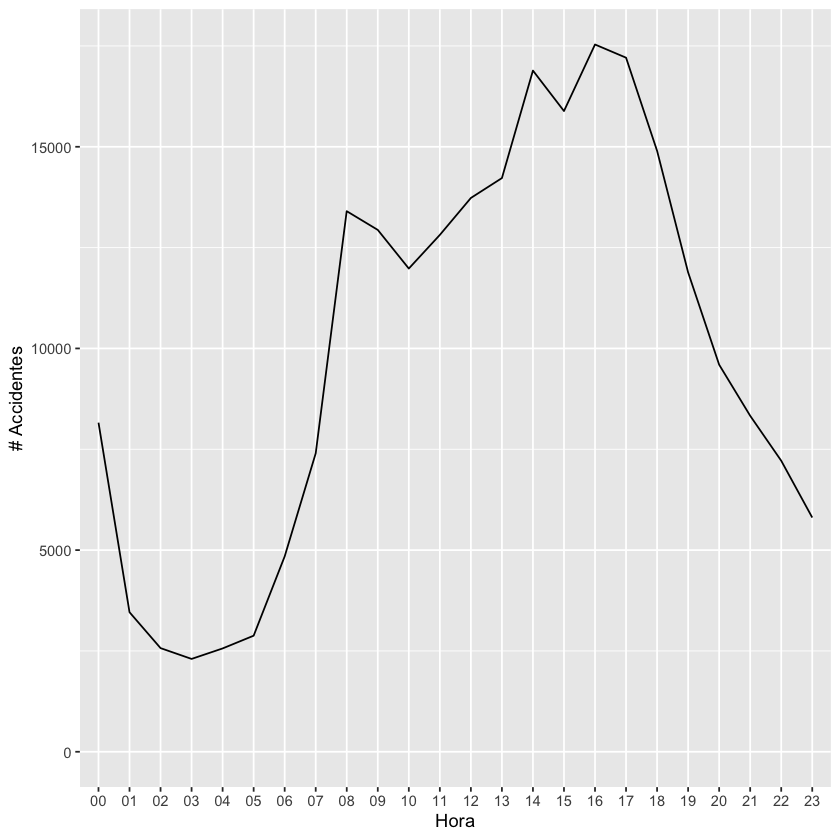
d %>% count(M.Y) %>% ggplot(aes(x=M.Y, y=n, group=1)) + geom\_line() + xlab("Mes") + ylab("# Accidentes") + scale\_y\_continuous(limits = c(0, NA))



## 2.2 ¿Existen patrones horarios en los accidentes?

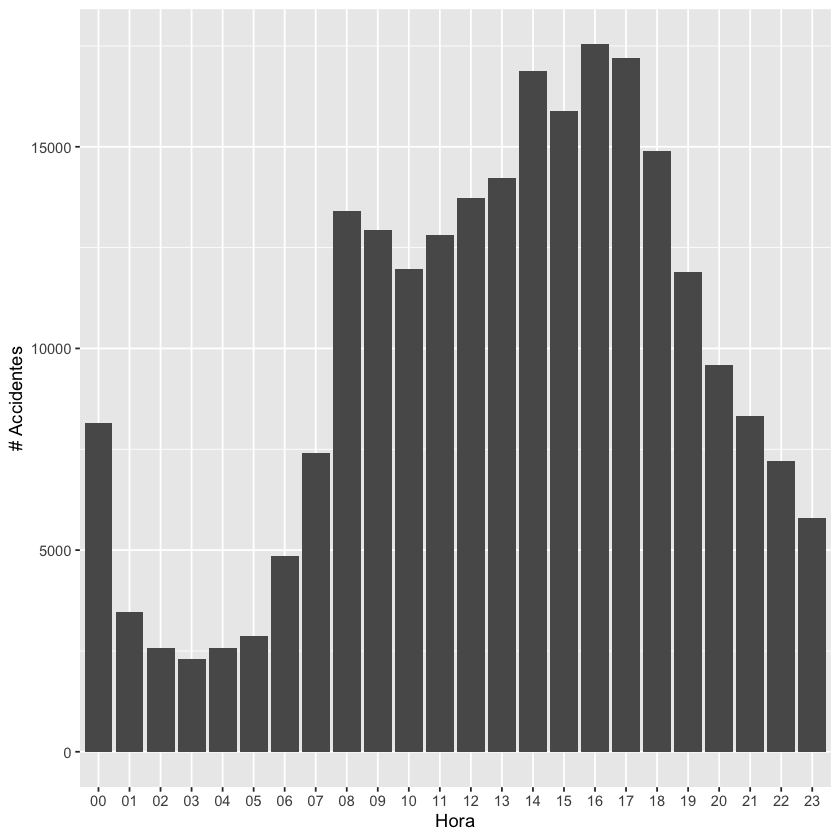
El patron de accidentes por hora esta determinado por la distribucion normal, lo que significa que al finalizar la tarde los accidentes se incrementan, las razones son variadas pero a simple vista es la hora en la que ciudadanos de NY, estan de retorno a sus lugares de residencia.

d %>% count(T) %>% ggplot(aes(x=T, y=n, group=1)) + geom\_line() + xlab("Hora") + ylab("# Accidentes") + scale\_y\_continuous(limits = c(0, NA))



Tambien se evidencia que dicha variable sigue una distribucion normal.

d %>% count(T) %>% ggplot(aes(x=T, y=n, group=1)) + geom\_bar(stat = "identity") + xlab("Hora") + ylab("# Accidentes")



## 2.3 ¿Existen patrones en los accidentes según el día de la semana?

Existen, algun dias se compartan de forma muy similar. Es el caso del Viernes, Jueeves y Martes, con ligera variacion los miercoles.

dw <- d %>% count(D)  
factor(dw$D, levels = c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday"))  
dwo <- dw[order(dw$D), ]  
dwo

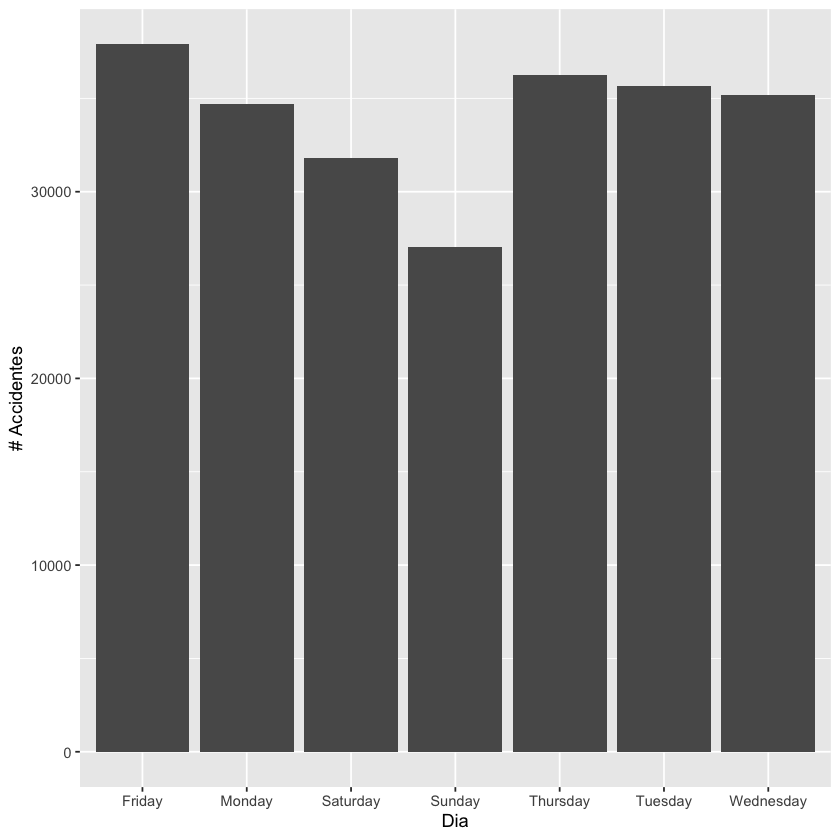
1. Friday
2. Monday
3. Saturday
4. Sunday
5. Thursday
6. Tuesday
7. Wednesday

**Levels**: 1. ‘Sunday’ 2. ‘Monday’ 3. ‘Tuesday’ 4. ‘Wednesday’ 5. ‘Thursday’ 6. ‘Friday’ 7. ‘Saturday’

A data.frame: 7 × 2

|  | D <fct> | n <int> |
| --- | --- | --- |
| 1 | Friday | 37886 |
| 2 | Monday | 34719 |
| 3 | Saturday | 31784 |
| 4 | Sunday | 27040 |
| 5 | Thursday | 36254 |
| 6 | Tuesday | 35651 |
| 7 | Wednesday | 35188 |

dwo %>% ggplot(aes(x=D, y=n)) + geom\_bar(stat = "identity") + xlab("Dia") + ylab("# Accidentes")



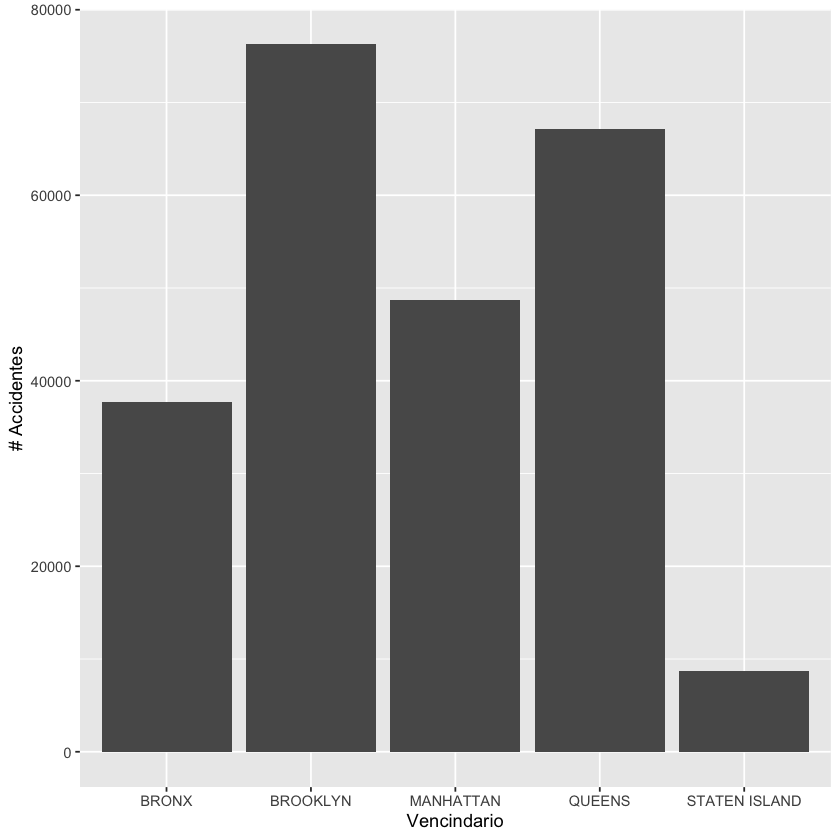
## 2.4 ¿Existen patrones accidentales por vecindario?

Su varianza se puede decir que es alta, a simple vista, pero si tuvieramos en cuenta que los tamanos de los vecindarios por area no son iguales. Para poder contestar esta pregunta se necesita incluir el area. El area fue obtenida del conjunto de datos borough\_data.csv. Borough data set contiene la siguiente informacion:

f <- read.csv2("./datasets/borough\_data.csv", sep = ",", header = TRUE)  
  
ft <- data.frame(  
 Columns = names(f),  
 Type = sapply(f, class),  
 Observations = sapply(f, length)  
)  
  
ft <- flextable(ft)  
ft <- theme\_vanilla(ft)  
ft

a flextable object.  
col\_keys: `Columns`, `Type`, `Observations`   
header has 1 row(s)   
body has 3 row(s)   
original dataset sample:   
 Columns Type Observations  
borough borough character 5  
population population integer 5  
area area character 5

d %>% count(BOROUGH) %>% ggplot(aes(x=BOROUGH, y=n, group=1)) + geom\_bar(stat = "identity") + xlab("Vencindario") + ylab("# Accidentes")



Ahora se puede asociar a cada vecindario un valor de area y con ello proceder a su visualizacion.

f$borough <- toupper(f$borough)  
f$borough[1] <- "BRONX"  
i <- d %>% count(BOROUGH)  
h <- left\_join(x = i, y = f, by = c('BOROUGH' = 'borough'))  
h$area <- as.numeric(h$area)  
a <- h %>% mutate(per\_area = n / area)

a

A data.frame: 5 × 5

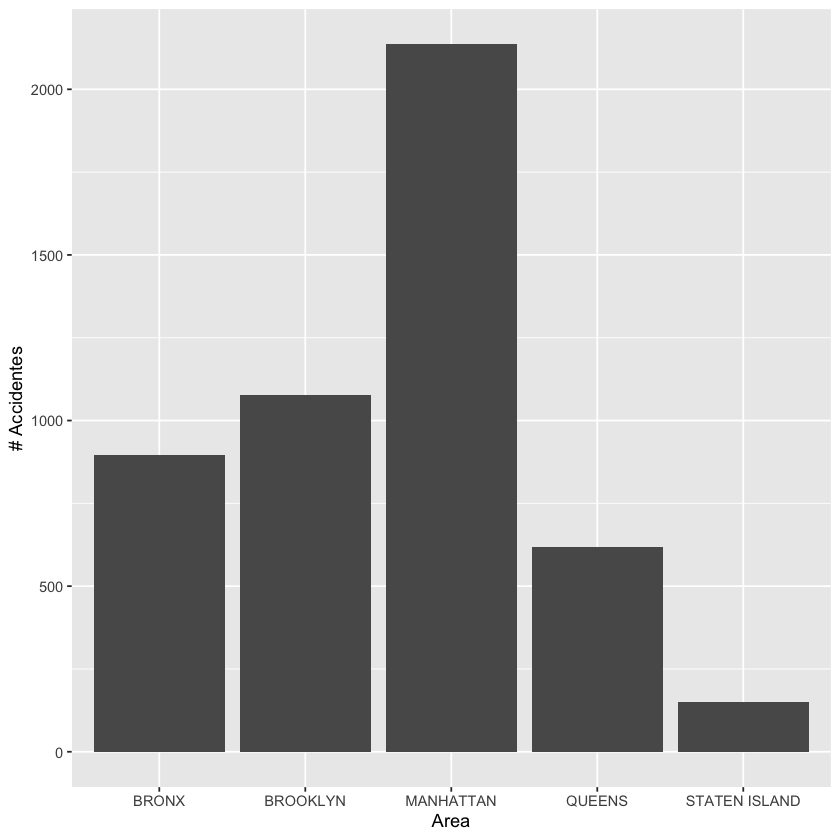
| BOROUGH <chr> | n <int> | population <int> | area <dbl> | per\_area <dbl> |
| --- | --- | --- | --- | --- |
| BRONX | 37709 | 1471160 | 42.10 | 895.7007 |
| BROOKLYN | 76253 | 2648771 | 70.82 | 1076.7156 |
| MANHATTAN | 48749 | 1664727 | 22.83 | 2135.3044 |
| QUEENS | 67120 | 2358582 | 108.53 | 618.4465 |
| STATEN ISLAND | 8691 | 479458 | 58.37 | 148.8950 |

h

A data.frame: 5 × 4

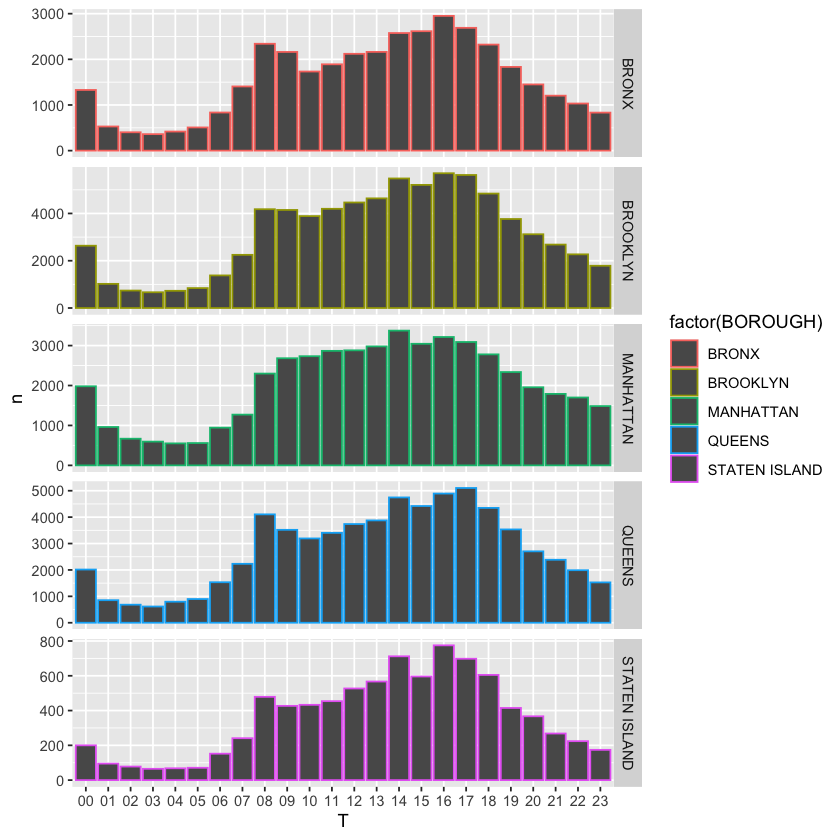
| BOROUGH <chr> | n <int> | population <int> | area <chr> |
| --- | --- | --- | --- |
| BRONX | 37709 | 1471160 | 42.1 |
| BROOKLYN | 76253 | 2648771 | 70.82 |
| MANHATTAN | 48749 | 1664727 | 22.83 |
| QUEENS | 67120 | 2358582 | 108.53 |
| STATEN ISLAND | 8691 | 479458 | 58.37 |

a %>% ggplot(aes(x=BOROUGH, y=per\_area, group=1)) + geom\_bar(stat = "identity") + xlab("Area") + ylab("# Accidentes")



Al relacionar la grafica de Accidentes x Area, se logra establecer que no existe una corelacion entra las dos variables. Por ende el nmumero de accidentes por area no muestra algun tipo de dependencia. Al comparar los distintos districtos de New York, vemos que su comportamiento es muy similar en relacion a su tamano, siguiendo una distribucion de probabilidad muy semejante a una normal.

d %>% count(BOROUGH, T) %>% ggplot(aes(x = T, y = n, colour = factor(BOROUGH))) + geom\_bar( stat = 'identity') + facet\_grid(vars(BOROUGH), scales = "free")



## 2.5 ¿Cuáles son las causas más comunes de los accidentes?

Son las leciones personales producto del accidente la causa de accidente mas comun. Seguido por los transeuntes y los ciclistas.

j <- d %>%  
 summarise(across(where(is.numeric), sum, na.rm = TRUE))  
  
jt <- flextable(j)  
jt <- theme\_vanilla(jt)  
jt

Warning message:  
“There was 1 warning in `summarise()`.  
ℹ In argument: `across(where(is.numeric), sum, na.rm = TRUE)`.  
Caused by warning:  
! The `...` argument of `across()` is deprecated as of dplyr 1.1.0.  
Supply arguments directly to `.fns` through an anonymous function instead.  
  
 # Previously  
 across(a:b, mean, na.rm = TRUE)  
  
 # Now  
 across(a:b, \(x) mean(x, na.rm = TRUE))”

a flextable object.  
col\_keys: `NUMBER.OF.PEDESTRIANS.INJURED`, `NUMBER.OF.PEDESTRIANS.KILLED`, `NUMBER.OF.CYCLIST.INJURED`, `NUMBER.OF.CYCLIST.KILLED`, `NUMBER.OF.MOTORIST.INJURED`, `NUMBER.OF.MOTORIST.KILLED`, `COLLISION\_ID`   
header has 1 row(s)   
body has 1 row(s)   
original dataset sample:   
 NUMBER.OF.PEDESTRIANS.INJURED NUMBER.OF.PEDESTRIANS.KILLED  
1 13005 127  
 NUMBER.OF.CYCLIST.INJURED NUMBER.OF.CYCLIST.KILLED NUMBER.OF.MOTORIST.INJURED  
1 5601 20 42475  
 NUMBER.OF.MOTORIST.KILLED COLLISION\_ID  
1 74 956006650080

# 3. Conclusiones

Si bien se lograron establecer algunos patrones como parte del analisis exploratorio, las politicas publicas deberian dirigirse a la educacion en los transeuntes especialmente en el area de manhatan dado que es el lugar en donde mas se presentan accidentes.

Lucas, Antonio, Francisco Alonso, Mireia Faus, and Arash Javadinejad. 2024. “The Role of News Media in Reducing Traffic Accidents.” *Societies* 14 (April): 56. <https://doi.org/10.3390/soc14050056>.