



Proyecto: Análisis de los datos de Airbnb como Data Science

(New York City)

Tabla de contenido

Presentación del proyecto.....	3
Propósito del proyecto y recogida de datos.....	3
Tratamiento de los datos.....	3
Análisis de los datos.....	5
Arboles de decisión.....	5
GLM	10
GAMs	12
CAUSAL IMPACT.....	13
XGBoost	14

Presentación del proyecto

Este proyecto forma parte de un trabajo de Trabajo Final de Máster (TFM) en el que las integrantes del mismo están trabajando. Este TFM consta de la realización de un recomendador de Airbnb con el objetivo de ayudar a los usuarios en la toma de decisiones sobre a qué ciudad viajar tomando como referencia aquellos interrogantes más inquietantes según distintos estudios ya realizados y documentados.

Dentro de estos interrogantes nos encontramos con: el tipo de alojamiento, el precio, la ubicación, el clima y la criminalidad.

Con el fin de, por un lado, ayudar a la toma de decisiones y, por el otro motivar a que la demanda aumente dando seguridad de este tipo de servicios de alojamientos no solo a gente que decide viajar a un modo low cost.

En el actual TFM se han escogido las 5 ciudades más visitas de EEUU, pero debido a la amplia extensión de ese proyecto, para esta ocasión vamos a presentar los datos incluyendo una única ciudad, New York City (NYC).

Propósito del proyecto y recogida de datos

El propósito de estudiar los datos, viene contemplado para ver qué dentro de los interrogantes arriba indicados tienen más peso, así como ver su distribución, su correlación y agrupación de los mismos.

Para ello, se han recogido los datos desde distintos datasheets mayormente obtenidos desde la página del gobierno de EEUU. Entre estos destacamos los datos de NYC de: crimen donde encontramos una gran variedad del tipo de datos relacionados con la criminalidad, datasheet del listado de apartamentos que tiene en dicha ciudad, datasheet del histórico de reserva de esos apartamentos y datasheet del clima de esa ciudad. A parte, se ha procedido a añadir unas bases de datos que hemos generado nosotras a través de la realización de web scraping en páginas de tripadvisor con el objetivo de calcular por barrios el número de hoteles, ocio y restaurantes que existen.

Tratamiento de los datos

Con esta gran variedad de datos, el tratamiento de los datos para conseguir una única base de datos con la que trabajar ha sido arduo y con una buena implicación en lo que a tiempo se refiere pues se han seguido los siguientes pasos:

- 1.) Observación y limpieza del tipo de datos, formatos, tipos de variables, variables interesantes que pueden servirnos y las que evitamos por si existiera un sobreajuste,...
- 2.) Con las bases de datos del listado de los alojamientos junto al histórico, realizamos una join cuya PrimaryKey es el id asociado a cada alojamiento.
- 3.) A esta nueva base de datos le realizamos una nueva join por fecha con la variable de crimen, no sin antes:
 - a. Se ha utilizado geojsons para poder determinar con exactitud el distrito y el barrio al que corresponde cada par de coordenadas que tenemos en los datasheets. El geojson contiene el rango de coordenadas de cada barrio y distrito. Desarrollando un código en Python, buscamos cada par de latitud y longitud de los datasheets en el geojson, y se añade nuevas columnas de barrio y distrito.

- b. Crear una variable para clasificar el tipo de crimen según su grado de severidad en: leve, medio, grave y se realiza una primera join para añadir esa clasificación a la base de datos de crimen.
- c. Dentro de esta nueva variable del tipo de crimen se procede a realizar un group by agrupando por fecha el número de delitos de ese día y por distrito. Es ese dato numérico el que se introduce en la base de datos conjunta a través de una nueva join.
- d. Como existen 24 horas al día y es complicado poder realizar observaciones con ello, creamos una nueva BBDD donde distinguimos la época en que se ha dado el acto delictivo según sea: madrugada, mañana, tarde y noche.

Como conclusión tenemos una BBDD donde se indica, la fecha del crimen, el tipo de crimen, severidad y hora del día; y son estos datos los que se añaden a la BBDD anterior.

4.) Obtenidos en una nueva base de datos las variables del listado, histórico y crimen, procedemos a introducir la variable tiempo, no sin antes realizar modificaciones:

- a. En la base de datos original observamos como existe una variable que corresponde al nivel de lluvia almacenado. Esta medida se encuentra en galones. Como la medida en sí no nos dice nada, se procede a generar una nueva base de datos acotándola por medidas, donde, según las medidas recogidas de agua, se le atribuye un margen del tipo de lluvia como puede ser: lluvia moderada, débil, tormenta,... . Se procede a realizar una join desde la medida en galones.
- b. Como es poco intuitivo observar la fecha del año para saber en qué estación estás y pocas conclusiones se pueden sacar si observamos las fechas como variable ya que hay 365 registros a lo largo del año, se procede a realizar otra base de datos donde se acoten las fechas por la estación del año. De esta forma nos resultará más fácil visualizar los datos y correlacionar las precipitaciones con la época del año.
- c. Se realiza una join con la base de datos anterior mediante la fecha.

Como conclusión obtenemos una base de datos con la estación del año, el día, las precipitaciones y las temperaturas, las cuales, a cuales esta última no hemos realizado ninguna alteración que anexionamos mediante join por fecha a la base de datos anterior que estamos creando de forma unificada.

5.) Ya unificados los datos del listado de alojamientos, el histórico, el crimen y las precipitaciones, procedemos a realizar otra join a través del barrio pero antes:

- a. A través de un group by, realizamos una agrupación de los tipos de servicio de ocio por barrio.
- b. Realizamos un count para que nos los cuente. Es este dato, al igual que en crimen, es el que utilizaremos para anexionar a la base de datos.
- c. Realizamos una join a través del barrio a la base de datos general donde nos indicará el total de restaurantes en la zona, de parques, supermercados,....

Finalizado todas estas modificaciones nos encontramos con un base de datos con los siguientes campos:

```
$ listing_id      : num [1:1168738] 2595 2595 2595 2595 2595 ...
$ date           : Date[1:1168738], format: "2009-11-21" "2009-12-05" "2009-12-10" "2010-04-09" ...
$ experiences_offered : chr [1:1168738] "No" "No" "No" "No" ...
$ host_since      : Date[1:1168738], format: "2008-09-09" "2008-09-09" "2008-09-09" "2008-09-09" ...
$ host_response_time : chr [1:1168738] "within a few hours" "within a few hours" "within a few hours" ...
$ host_total_listings_count: num [1:1168738] 6 6 6 6 6 6 6 6 ...
$ district        : chr [1:1168738] "Manhattan" "Manhattan" "Manhattan" "Manhattan" ...
$ is_location_exact : chr [1:1168738] "No" "No" "No" "No" ...
$ property_type    : chr [1:1168738] "Apartment" "Apartment" "Apartment" "Apartment" ...
$ room_type        : chr [1:1168738] "Entire home/apt" "Entire home/apt" "Entire home/apt" "Entire home/apt" ...
```

```
$ bedrooms      : num [1:1168738] 0 0 0 0 0 0 0 0 ...
$ beds         : num [1:1168738] 1 1 1 1 1 1 1 1 ...
$ price        : num [1:1168738] 210 210 210 210 210 210 210 210 ...
$ security_deposit : num [1:1168738] 154 154 154 154 154 154 154 154 ...
$ cleaning_fee  : num [1:1168738] 218 218 218 218 218 218 218 218 ...
$ minimum_nights : num [1:1168738] 3 3 3 3 3 3 3 3 ...
$ calendar_updated : chr [1:1168738] "3 months ago" "3 months ago" "3 months ago" "3 months ago" ...
$ last_review   : Date[1:1168738], format: "2019-11-04" "2019-11-04" "2019-11-04" "2019-11-04" ...
$ review_scores_accuracy : num [1:1168738] 9 9 9 9 9 9 9 9 ...
$ instant_bookable : logi [1:1168738] FALSE FALSE FALSE FALSE ...
$ cancellation_policy : chr [1:1168738] "strict_14_with_grace_period" "strict_14_with_grace_period" ...
$ state        : chr [1:1168738] "New York City" "New York City" "New York City" "New York City" ...
$ TMAX         : num [1:1168738] 14.44 8.33 8.89 16.67 23.33 ...
$ TMIN         : num [1:1168738] 6.67 1.11 -2.22 7.78 13.89 ...
$ PRCP_Type    : chr [1:1168738] "No rain" "Heavy rain" "No rain" "Moderate rain" ...
$ Season       : chr [1:1168738] "Autumn" "Winter" "Winter" "Spring" ...
$ Crime_Minor  : num [1:1168738] 192 171 174 196 154 153 199 146 149 183 ...
$ Crime_Misdemeanor : num [1:1168738] 123 109 131 126 116 103 107 89 85 119 ...
$ Crime_Serious : num [1:1168738] 42 57 46 64 52 46 61 50 41 51 ...
$ School       : num [1:1168738] 793 793 793 793 793 793 793 793 793 ...
$ Restaurant   : num [1:1168738] 10625 10625 10625 10625 10625 ...
$ Park         : num [1:1168738] 31 31 31 31 31 31 31 31 31 ...
$ Shopping_mall : num [1:1168738] 209 209 209 209 209 209 209 209 ...
```

Generada esta base de datos observamos que nos encontramos con más de un millón de registros, cuya fecha comienza en 2009, un año después del nacimiento de esta compañía. Debido a esta gran cantidad de datos, procedemos a realizar un estudio de los datos a partir del 2015. Se ha escogido el 2015 como año de partida debido a que el despegue de esta compañía fue en este año.

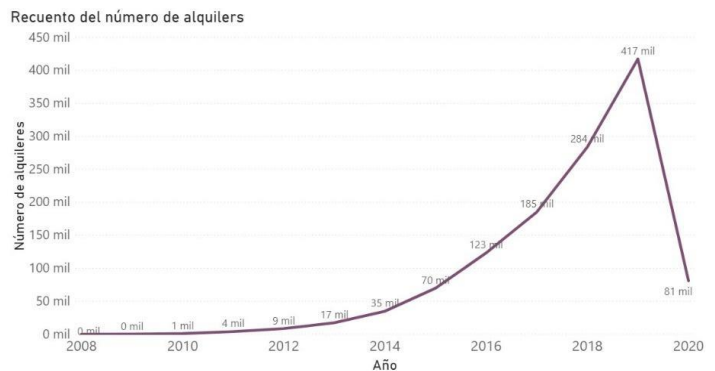


Ilustración 1. Recuento del número de alquileres

Análisis de los datos

Como primera opción para este proyecto hemos decidido ver el tipo de clasificaciones que podemos tener en los datos para tener un mayor conocimiento de los datos que tenemos.

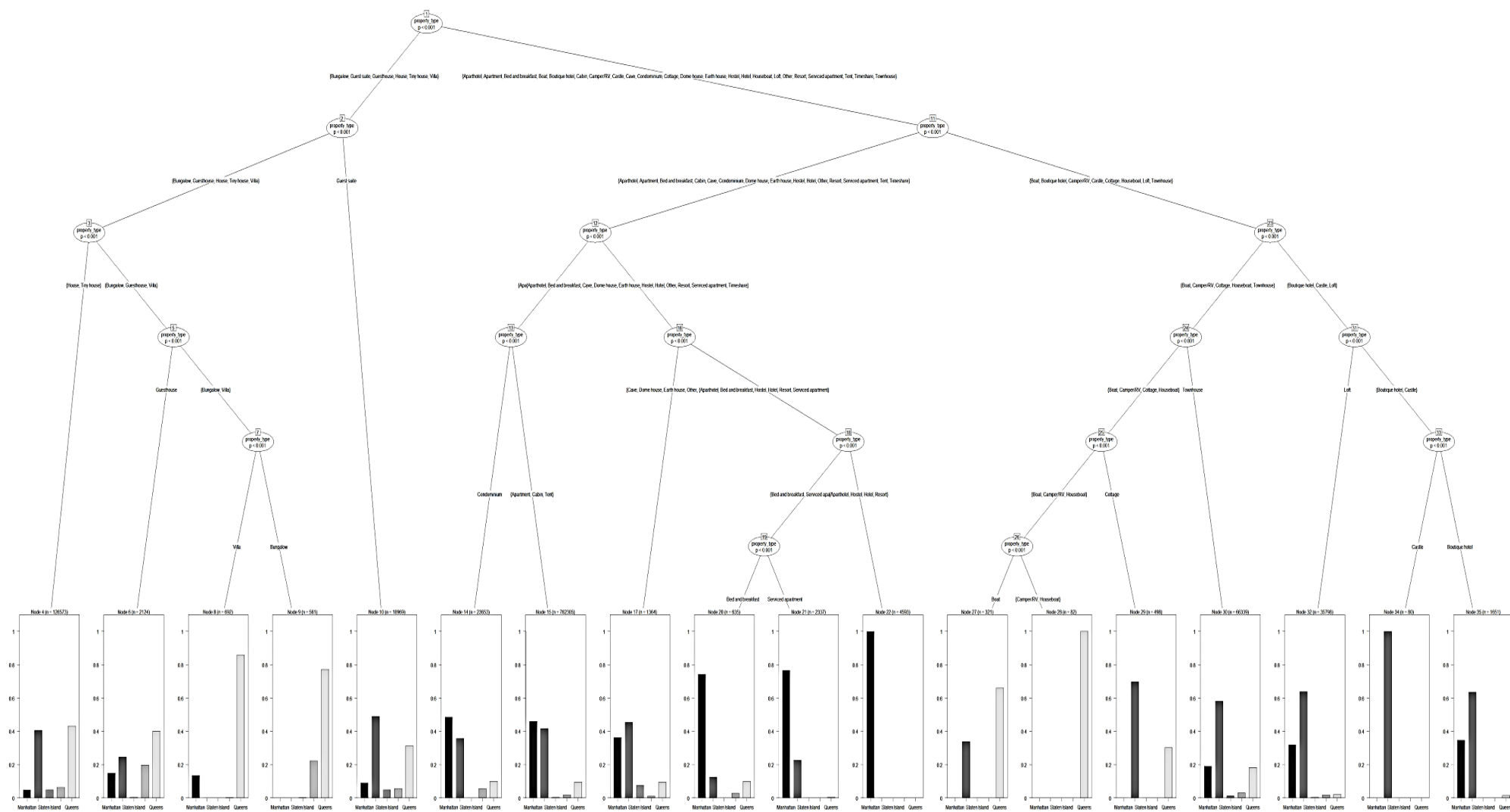
Arboles de decisión

- a.) Como primera opción y, observando el tipo de datos que tenemos procedemos a observar si se puede hacer diferenciación con el tipo de vivienda y el distrito en el que se encuentra.

Lo primero que hacemos es ver el tipo de distritos que hay y el número de alquileres por distrito. Obtenemos la siguiente tabla:

Manhattan	Brooklyn	Staten Island	Bronx	Queens
403016	452606	12629	27979	152345

Ilustración 2. Relación entre tipo de vivienda y distrito



En la imagen podemos comprobar que existen dos claras diferenciaciones, una son: Bungalow, Guest suite, Guesthouse, House, Tiny house y Villa, y por el otro, el resto. Dentro de esta diferenciación, observamos que:

- Bungalow: se distribuyen en dos distritos, Queens con casi un 80% y Bronx con casi el 20%
- Guest Suite: la distribución es de un 50% en State Island y un 30% en Queens, el resto se reparte de forma más o menos equitativa.
- Guesthouse: aunque Queens tiene algo más del 40%, del resto, salvo Staten Island que apenas se dan, del resto de distritos está repartido.
- House y Tiny house: se valora con un 40% en Staten Island y Queens otro 40%, repartidos el resto de distritos
- Villa: encontramos con aproximadamente un 85% que este tipo de vivienda se da en Queens, un 15% en Manhattan y el resto en Bronx.

Con el resto del tipo de viviendas, se destaca lo siguientes:

- Que los distritos donde mayormente se encuentran los hoteles, apartahoteles, resorts y servicio de alojamiento y desayuno se dan en Manhattan seguido de Brooklyn.
- Que el alojamiento en barcos o propiedades similares y furgonetas se dan en Queens y Brooklyn.
- Que en el distrito de Brooklyn podemos encontrar además, diversidad de tipos de alojamientos como castillo, hoteles boutique, lofts, granjas,... prevaleciendo este distrito como el más numeroso entre todos ellos, llegando en algunos tipos el único que estos ofrece.

b.) Por otro lado, quisimos saber si existía alguna diferenciación entre las políticas de cancelación de las reservas y los distritos. Decidimos observar este patrón ya que sería interesante estudiar este dato con respecto a ayudar a los viajeros en su toma de decisiones sobre el vecindario en el que instalarse.

En la gráfica observamos que aunque existe división entre las políticas de flexible, moderado y estricto, el mayor número se dan en los distritos de Manhattan, Brooklyn y Queens.

Si observamos esta relación con el número de alojamientos, observamos que estos tres distritos son los que más número de alojamientos disponibles se dan, con lo cual, podemos obtener diversas políticas entre los distintos alojamientos. En este caso tendríamos que mirar el tipo alojamiento que, como se demuestra en la gráfica posterior, dependiendo del tipo de alojamiento se diferencia una política de cancelación u otra. La decisión del tipo de alojamiento es algo más particular a la hora de la elección del alquiler, por lo que no nos adentraremos en detalle a comentar estos resultados, aunque se muestran en la siguiente ilustración.

Ilustración 3. Relación política de cancelación con distrito

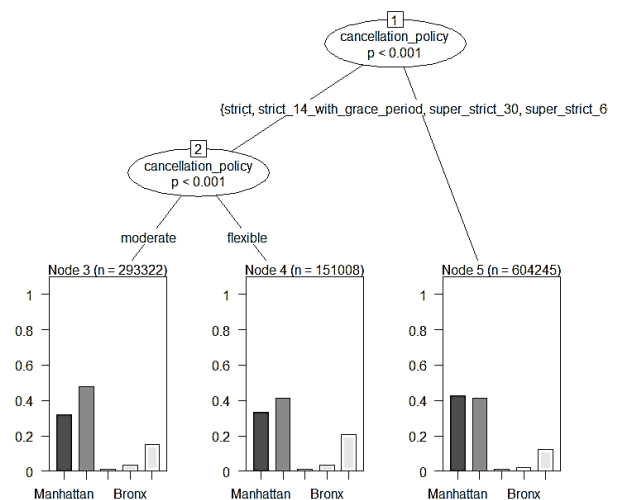
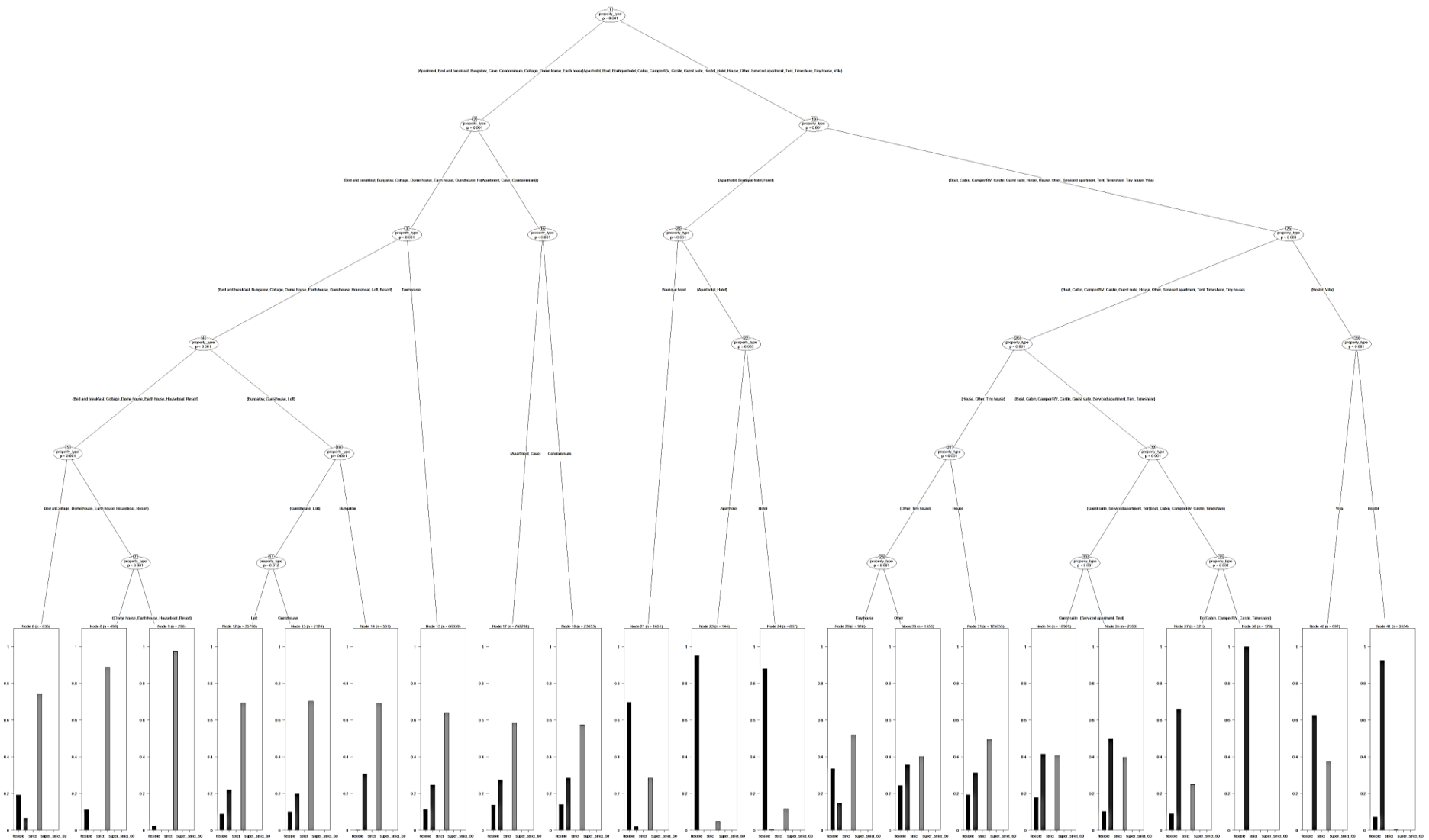


Ilustración 4. Relación alojamientos con distrito



c.) Ahora vamos a observar si los factores externos afectan a los alquileres. En esta ocasión, observamos el tipo de registros que tenemos y nos encontramos con la siguiente tabla:

Autumn : 315831

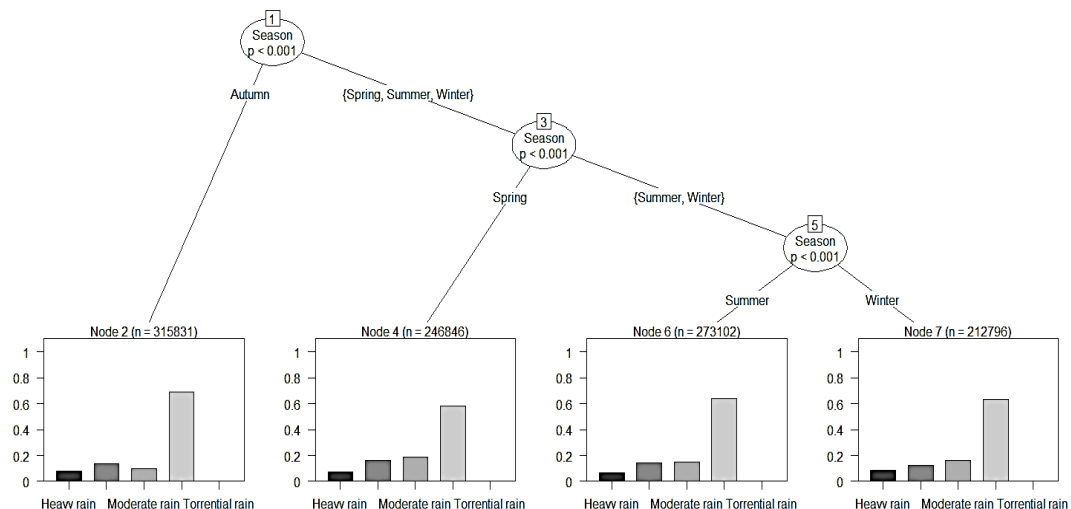
Summer : 273102

Spring : 246846

Winter : 212796

A todo ello, observamos las precipitaciones que se dan a lo largo del año. En ellas observamos que no existe una gran diferenciación entre los distintos meses.

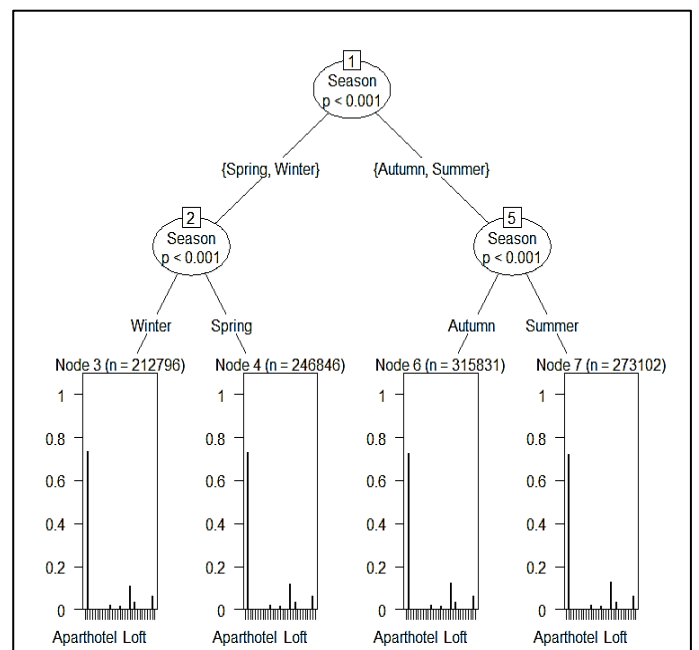
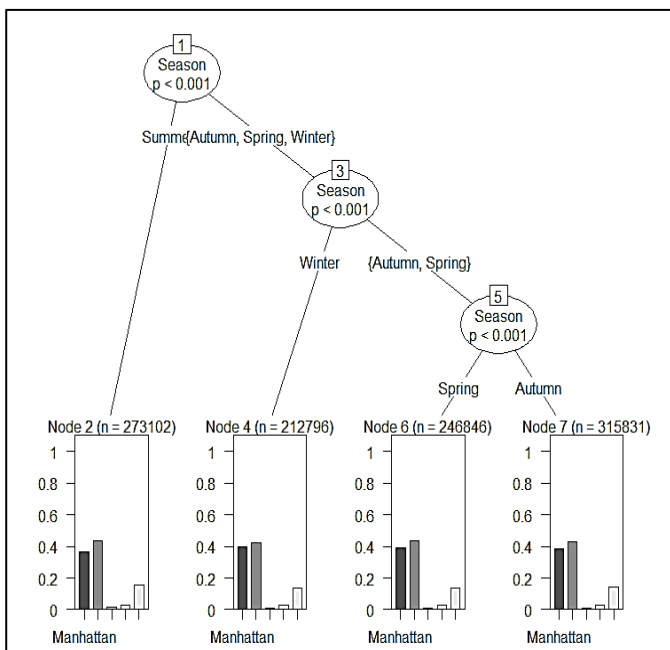
Ilustración 5. Relación estación del año con precipitaciones



En las gráficas siguientes, observamos la misma similitud de datos que los mostrados en la tabla anterior. En ella observamos que apenas hay diferenciación con respecto a la hora de escoger Nueva York como destino turístico a lo largo del año, teniendo esta ciudad continuamente turistas y estas se distribuyen de igual manera entre los distintos distritos.

Ilustración 7. Relación estación del año con distrito.

Ilustración 6. Relación estación del año con tipo de alojamiento.



GLM

A continuación, pasamos a estudiar la distribución que presenta el número de alquileres a lo largo de estos años y medir el posible impacto que haya podido causar el COVID-19.

Para ello, pasaremos a modelar los datos con un Modelo Lineal Generalizado o GLM, concretamente con una distribución de Poisson. Este tipo de distribuciones se utilizan cuando las variables de interés son números enteros y mayores a 0, y son eventos independientes entre sí.

En primer lugar, agrupamos los alquileres por fecha y dibujamos los datos obteniendo la siguiente gráfica.

Se puede observar que existe una cierta estacionalidad. En los meses de primavera y verano el número de alquileres se incrementa, coincidiendo con las fechas en las que se suele tener vacaciones y el clima es más favorable.

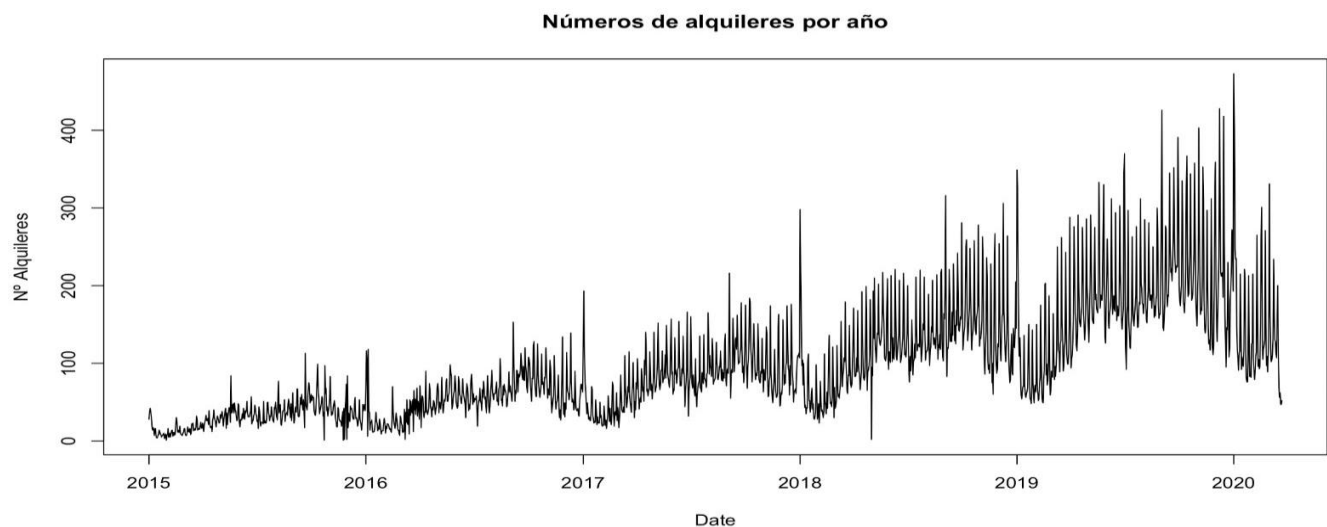


Ilustración 8. Número de alquileres por año

Dado que existen muchos datos, se decide filtrar el dataset para trabajar solamente con los últimos 4 años.

Con el primer modelo, queremos ver si el número de alquileres aumenta en el tiempo, por lo que se decide crear la variable día.

Vemos que, efectivamente, cada día se incrementa $7.585e-04$ el número de alquileres.

```
> summary(modelo01)

Call:
glm(formula = Registros ~ dia, family = poisson, data = airbnb3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-11.8835  -2.3966  -0.3132   1.6227  11.9673

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.402e+00  9.235e-03  368.43  <2e-16 ***
dia          7.585e-04  1.176e-05   64.51  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ilustración 9. Script de R del modelo anterior

Seguidamente, hacemos un predict y dibujamos los datos. Confirmamos que la tendencia de alquileres es creciente.

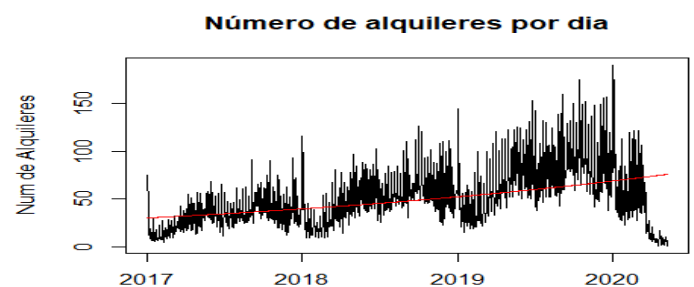


Ilustración 10. Número de alquileres por día

Para mejorar el modelo, se decide introducir una nueva variable, en este caso, el día de la semana.

Tras la obtención del summary, vemos que respecto al Domingo, el número de alquileres disminuye el resto de días de la semana (los coeficientes presentan signo negativo).

Ilustración 12. Script de pantalla R del modelo

```
> summary(modelo02)

Call:
glm(formula = Registros ~ dia + dow, family = poisson, data = tmp)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-27.709   -3.436    0.445    3.383   27.812

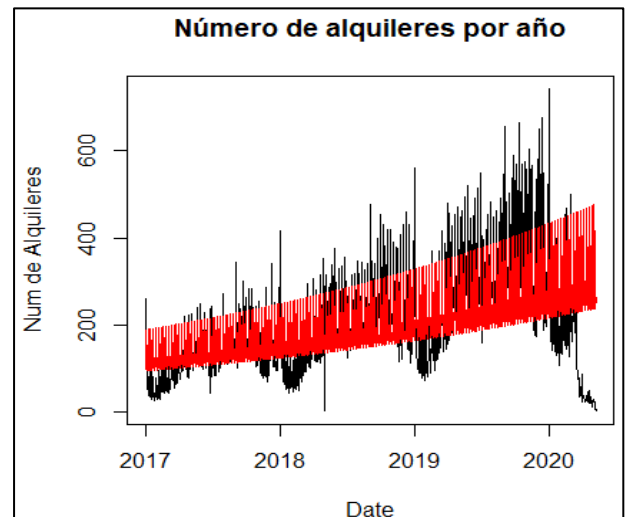
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.235e+00  5.963e-03  877.92  <2e-16 ***
dia          7.628e-04  5.880e-06  129.74  <2e-16 ***
dowjueves    -6.978e-01  7.439e-03  -93.80  <2e-16 ***
dowlunes     -2.736e-01  6.529e-03  -41.90  <2e-16 ***
dowmartes    -6.269e-01  7.271e-03  -86.22  <2e-16 ***
dowmiércoles -6.906e-01  7.423e-03  -93.04  <2e-16 ***
dowsábado    -4.899e-01  6.974e-03  -70.25  <2e-16 ***
dowviernes   -6.032e-01  7.210e-03  -83.66  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 80089  on 1223  degrees of freedom
Residual deviance: 47015  on 1216  degrees of freedom
AIC: 55510

Number of Fisher Scoring iterations: 5
```

Ilustración 11. Plot del script anterior



Pasamos a introducir dos variables, seno y coseno, para ver si mejora nuestro modelo. De esta forma introducimos una oscilación de 365 días.

Como se puede observar, el modelo ajusta mucho mejor. Aun así, las oscilaciones de los datos no coinciden completamente con el coseno. Por este motivo, se decide probar con un **modelo GAM** ya que tenemos conocimientos a priori de los datos y nos permite crear modelos más flexibles que puedan ajustarse mejor a nuestros datos.

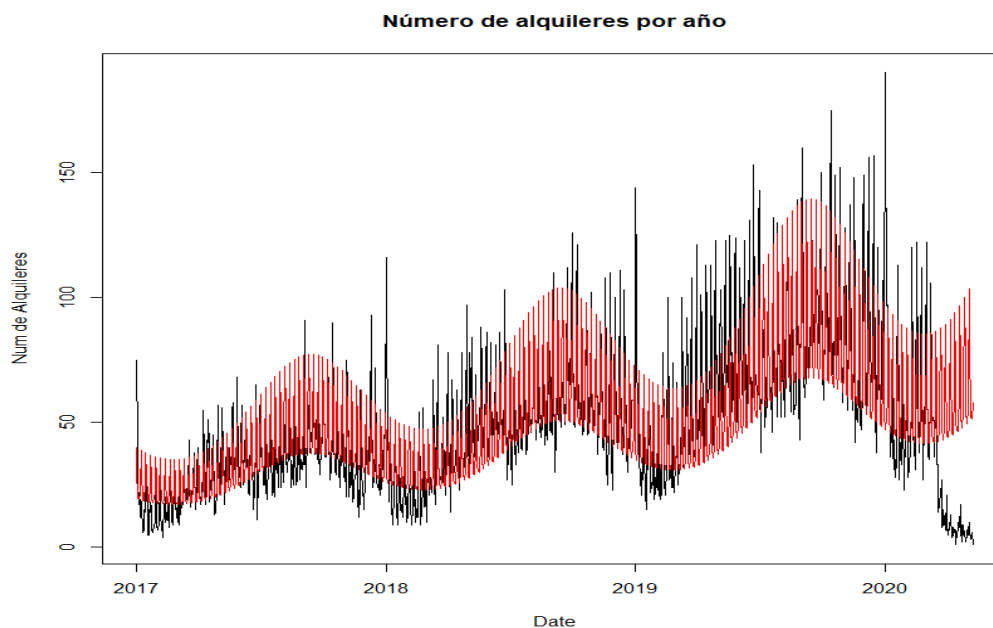


Ilustración 13. Plot del número de alquileres con seno y coseno

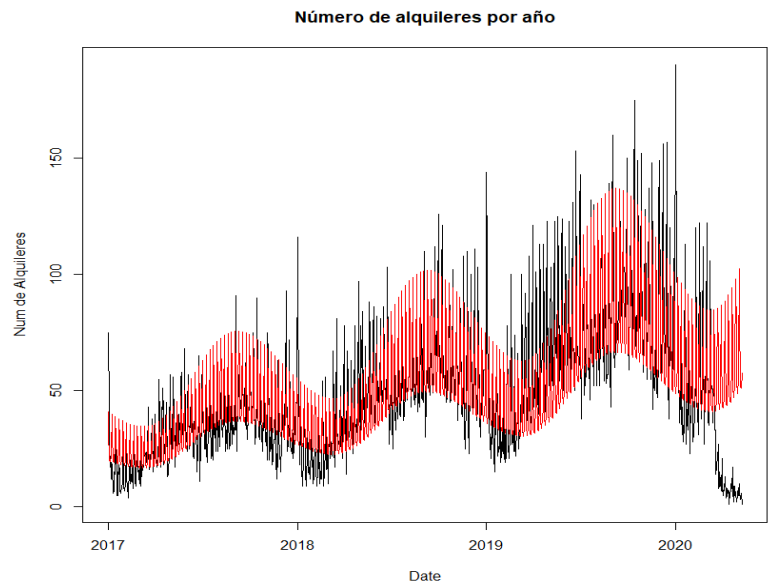
GAMs

Comenzamos creando una nueva variable que contenga el día del año.

Configuramos un spline con la nueva variable, que permita que su valor se repita cíclicamente, sin discontinuidades. Se define $k=5$, ya que vemos que está bastante equilibrado. A mayor número de nodos, mayor flexibilidad del modelo, pero se corre el riesgo de sobreajustar el modelo por lo que se mantiene ese número de nodos.

Se puede observar que el modelo se ajusta ligeramente mejor a las oscilaciones de los datos, gracias al spline cíclico.

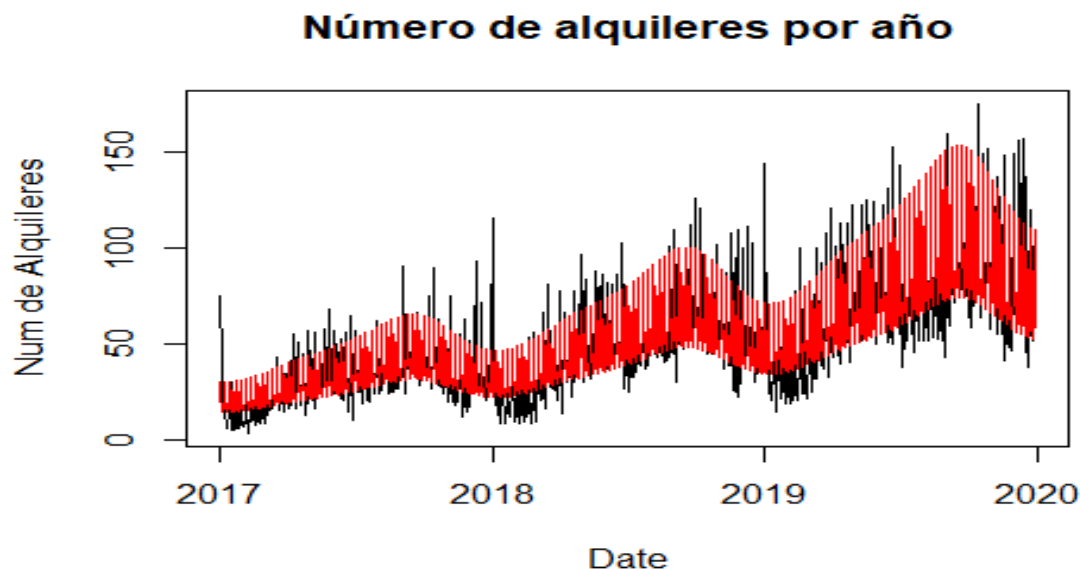
Ilustración 14. Plot con modelo GAMs



Dado el impacto que ha producido el COVID-19 este año en el turismo, creemos que es importante hacer la predicción sin tener en cuenta este periodo. Por ello, se genera un nuevo modelo filtrando los datos hasta principios de este año.

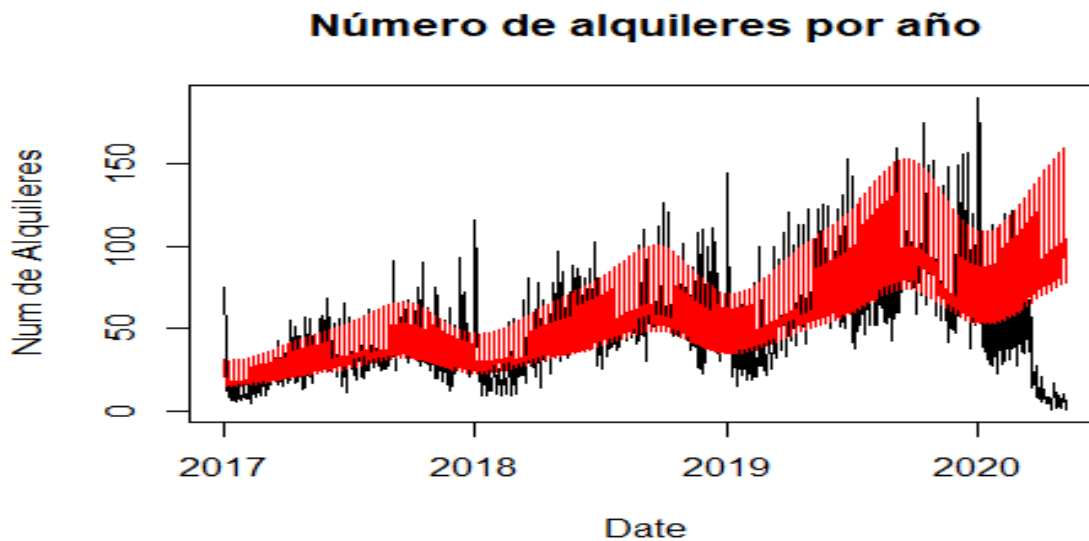
Como se puede ver en la gráfica resultante, el modelo se ajusta mucho mejor tras la eliminación del periodo COVID19.

Ilustración 15. Plot modelo GAMs 2



Finalmente, aplicamos esta nueva predicción a todo el conjunto de datos. Se puede observar la gran diferencia que existe entre la predicción (cómo serían los datos si no hubiera surgido la pandemia) y los datos reales en los que se observa la gran caída producida en el número de alquileres.

Ilustración 16. Plot modelo GAMs 3



CAUSAL IMPACT

Para medir qué impacto ha tenido el COVID-19 en el número de alquileres de apartamentos en Airbnb, procedemos a utilizar el paquete de **CausalImpact**.

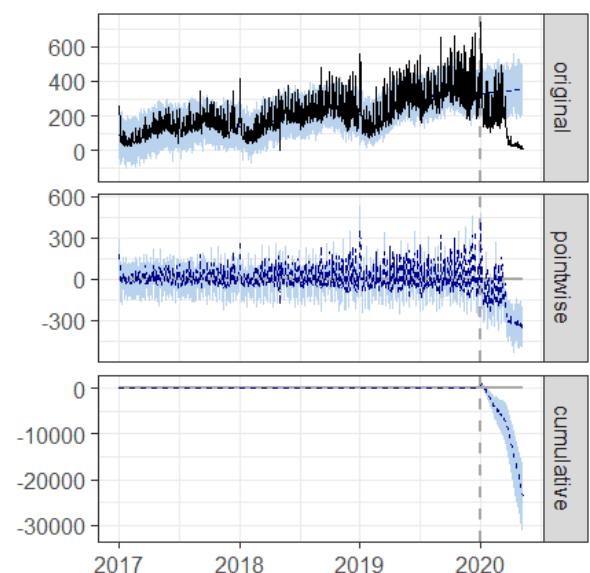
Analysis report {CausalImpact} viviendas

El efecto negativo observado durante el período de intervención es estadísticamente significativo. La probabilidad de obtener este efecto por casualidad es muy pequeña (probabilidad de área de cola unilateral bayesiana $p = 0.014$).

Se observa una caída significativa en el número de alquileres tras 2020, esto es consecuencia del impacto negativo que la pandemia de la Covid-19 está teniendo en Airbnb. No se está realizando reservas de hoteles debido al confinamiento y esto se visualiza con la caída del número de alquileres.

Por otro lado, se ha decidido medir qué impacto había causado el COVID-19 en el número de crímenes cometidos en la ciudad de Nueva York.

Ilustración 17. Analysis report {CausalImpact} alquileres



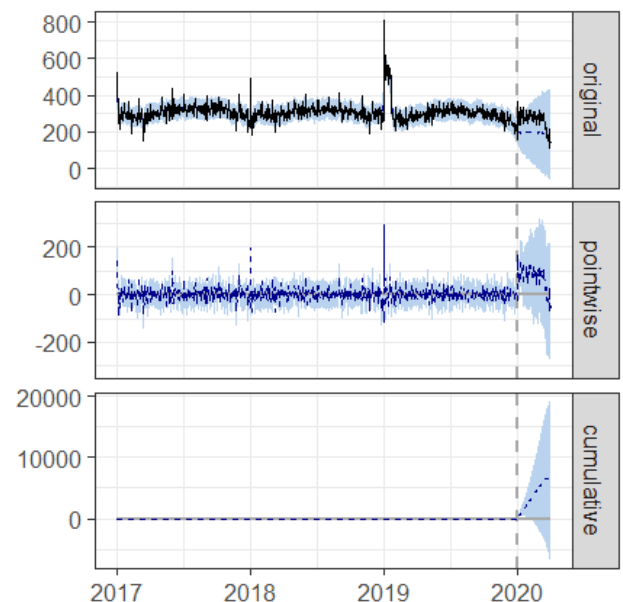
Analysis report {CausalImpact} Crimen

Una vez realizado el análisis, obtenemos que el impacto no era significativo ($p = 0.161$). Esto puede ser debido a que solo disponíamos de datos hasta finales de marzo, por lo que no era posible todavía ver el efecto.

Como se puede observar en las gráficas el Covid ha tenido un gran impacto, ya que si esto no hubiera sucedido, los alquileres no hubieran bajado. Esta gran bajada hace que la empresa Airbnb sufra una elevada pérdida de dinero en el año 2020, pudiendo afectar a los siguientes años.

Como dato curioso que, desvía el pronóstico, observamos cómo se da un pico a principios de 2019. Este dato nos ha llamado la atención, por lo que hemos procedido a realizar una investigación a través de los buscadores. Según el medio de comunicación el Diario, el 21 de febrero de 2019 nos da la noticia de que el crimen en la ciudad de Nueva York aumenta un 55% , dándose un incremento del 400% frente al 2018 de los asesinatos en la zona de Brooklyn. Este dato, verifica que nuestro modelo es correcto y que ese desvío, está justificado.¹

Ilustración 18. Analysis report {CausalImpact} Crimen



XGBoost

XGBoost parte del árbol de decisiones para clasificar o pronosticar sobre una variable objetivo (y). Aunque utilizan estos clasificadores débiles, XGBoost potencia los resultados debido al procesamiento secuencial de los datos con una función de pérdida o coste, la cual, minimiza el error de iteración tras iteración, haciéndolo de esta manera, un pronosticador fuerte.

Tras los análisis de los árboles de decisión, hemos decidido aplicar XGBoost para estudiar nuestras variables objetivo: precio del hotel. El objetivo es conocer qué variables son las más importantes para nuestra variable “y”.

Los parámetros ajustados para la rejilla son nround y eta. Se ha asignado diferentes números de iteraciones: 400, 1000 y 2400. Para el tamaño de salto en cada modelo nuevo se probado los valores 0.05 y 0.12.

¹ <https://eldiariiony.com/2019/02/21/aumentan-55-los-asesinatos-en-nyc-en-lo-que-va-del-2019/>

Ilustración 19. Script 1. Variable de estudio: precio. Valores $nround = 1000$ y $eta = 0.12$

```
> # importancia de variables:
> varImp(modelo)
xgbTree variable importance
```

only 20 most important variables shown (out of 32)

	Overall
room_type	100.000
listing_id	71.135
cleaning_fee	52.769
host_since	51.385
last_review	31.002
beds	24.255
security_deposit	22.508
bedrooms	19.684
host_total_listings_count	16.807
calendar_updated	16.335
School	13.793
minimum_nights	13.570
property_type	12.724
district	10.534
Park	9.575
host_response_time	6.360
Restaurant	5.046
cancellation_policy	4.096
review_scores_accuracy	3.983
instant_bookable	3.547

Ilustración 20. Script 2. Variable de estudio: precio. $nround = 2400$ y $eta = 0.05$

```
> varImp(modelo)
xgbTree variable importance
```

only 20 most important variables shown (out of 32)

	Overall
room_type	100.000
listing_id	60.404
host_since	43.723
cleaning_fee	35.509
last_review	26.483
bedrooms	17.823
beds	15.711
security_deposit	15.310
host_total_listings_count	14.506
School	14.339
calendar_updated	12.422
property_type	10.200
Restaurant	9.639
minimum_nights	9.637
host_response_time	5.691
district	4.615
review_scores_accuracy	3.821
cancellation_policy	2.958
Shopping.mall	2.450
instant_bookable	2.412

Tras ver estos dos resultados del precio, nos hemos dado cuenta del peso que está teniendo la variable `listing_id`. Esta variable es el id de los hoteles, para mejorar el modelo vamos a eliminar esta variable y repetir el modelo. El nuevo resultado obtenido se muestra en la siguiente imagen.

Se observa que las variables más importantes para el precio son: tipo de habitación, desde cuándo es el hotel, coste de limpieza, último review, si hay escuelas por el distrito, número de camas, cantidad de depósito, entre otros. Esta conclusión resulta coherente, ya que las características de la habitación y los costes de depósito y limpieza suelen ser los parámetros que más se tiene en cuenta a la hora de realizar una reserva de hotel.

Ilustración 21. Script 3. Variable de estudio: precio. $nround = 2400$ y $eta = 0.05$

```
> varImp(modelo)
xgbTree variable importance
```

only 20 most important variables shown (out of 31)

	Overall
room_type	100.000
host_since	36.224
cleaning_fee	33.864
last_review	18.453
School	15.442
beds	15.236
security_deposit	13.863
bedrooms	13.711
host_total_listings_count	12.271
calendar_updated	9.855
property_type	9.140
minimum_nights	7.814
Park	6.237
Restaurant	3.985
host_response_time	3.603
district	3.316
review_scores_accuracy	3.262
cancellation_policy	2.596
Shopping.mall	2.210
instant_bookable	1.739