CASO 1: NEW YORK CITY TAXI

1. Análisis cuantitativo

1.1 Primer examen preliminar del dataset. ¿Qué parámetros hay en el dataset? ¿Cuál es su significado? ¿Existen valores aparentemente incorrectos?

Field Name	Description
VendorID	A code indicating the LPEP provider that provided the record. **1= Creative Mobile Technologies, LLC; 2= VeriFone Inc.**
lpep_pickup_datetime	The date and time when the meter was engaged.
lpep_dropoff_datetime	The date and time when the meter was disengaged.
Passenger_count	The number of passengers in the vehicle. This is a driver-entered value.
Trip_distance	The elapsed trip distance in miles reported by the taximeter.
PULocationID	TLC Taxi Zone in which the taximeter was engaged
DOLocationID	TLC Taxi Zone in which the taximeter was disengaged
RateCodeID	The final rate code in effect at the end of the trip. **1= Standard rate** **2=JFK** **3=Newark** **4=Nassau or Westchester** **5=Negotiated fare** **6=Group ride**
Store_and_fwd_flag	This flag indicates whether the trip record was held in vehicle memory before sending to the vendor, aka "store and forward," because the vehicle did not have a connection to the server. **Y= store and forward trip** **N= not a store and forward trip**
Payment_type	A numeric code signifying how the passenger paid for the trip. **1= Credit card** **2= Cash** **3= No charge** **4= Dispute** **5= Unknown** **6= Voided trip**
Fare_amount	The time-and-distance fare calculated by the meter.
Extra	Miscellaneous extras and surcharges. Currently, this only includes the $0.50 and $1 rush hour and overnight charges.
MTA_tax	$0.50 MTA tax that is automatically triggered based on the metered rate in use.
Improvement_surcharge	$0.30 improvement surcharge assessed on hailed trips at the flag drop. The improvement surcharge began being levied in 2015.
Tip_amount	Tip amount – This field is automatically populated for credit card tips. Cash tips are not included.
Tolls_amount	Total amount of all tolls paid in trip.
Total_amount	The total amount charged to passengers. Does not include cash tips.

Figura 1 Variables dataset NYC Green

Field Name	Description
VendorID	A code indicating the TPEP provider that provided the record. **1= Creative Mobile Technologies, LLC; 2= VeriFone Inc.**
tpep_pickup_datetime	The date and time when the meter was engaged.
tpep_dropoff_datetime	The date and time when the meter was disengaged.
Passenger_count	The number of passengers in the vehicle. This is a driver-entered value.
Trip_distance	The elapsed trip distance in miles reported by the taximeter.
PULocationID	TLC Taxi Zone in which the taximeter was engaged
DOLocationID	TLC Taxi Zone in which the taximeter was disengaged
RateCodeID	The final rate code in effect at the end of the trip. **1= Standard rate** **2=JFK** **3=Newark** **4=Nassau or Westchester** **5=Negotiated fare** **6=Group ride**
Store_and_fwd_flag	This flag indicates whether the trip record was held in vehicle memory before sending to the vendor, aka "store and forward," because the vehicle did not have a connection to the server. **Y= store and forward trip** **N= not a store and forward trip**
Payment_type	A numeric code signifying how the passenger paid for the trip. **1= Credit card** **2= Cash** **3= No charge** **4= Dispute** **5= Unknown** **6= Voided trip**
Fare_amount	The time-and-distance fare calculated by the meter.
Extra	Miscellaneous extras and surcharges. Currently, this only includes the $0.50 and $1 rush hour and overnight charges.
MTA_tax	$0.50 MTA tax that is automatically triggered based on the metered rate in use.
Improvement_surcharge	$0.30 improvement surcharge assessed trips at the flag drop. The improvement surcharge began being levied in 2015.
Tip_amount	Tip amount – This field is automatically populated for credit card tips. Cash tips are not included.
Tolls_amount	Total amount of all tolls paid in trip.
Total_amount	The total amount charged to passengers. Does not include cash tips.

Figura 2 Variables dataset NYC Yellow

Sí, hay valores incorrectos. Se realiza una limpieza de los mismos para el posterior análisis del dataset. Dicha limpieza se puede observar en el código adjunto.

1.2 Empezamos por visualizar el dataset. Haced un plot de los puntos de recog ida y otro con los puntos de llegada del dataset de los cuatro ficheros y ext rae conclusiones preliminares. ¿Se aprecian diferencias entre los puntos d e los Yellow Cabs y los de los Green Cabs? ¿Se aprecian diferencias entre u n mes y otro?

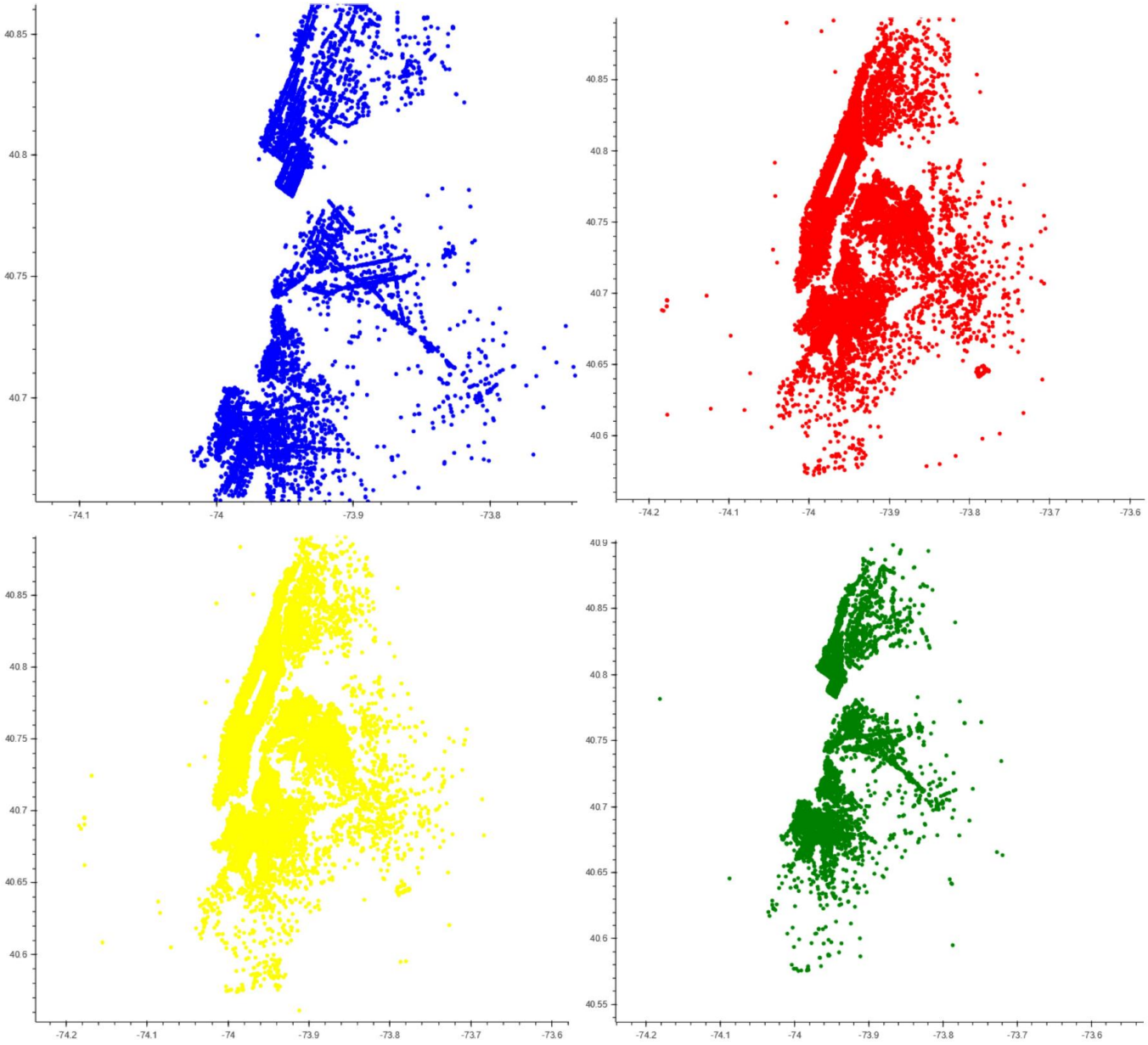

Figura 3 Plot de los puntos de recogida y llegada NYC Green

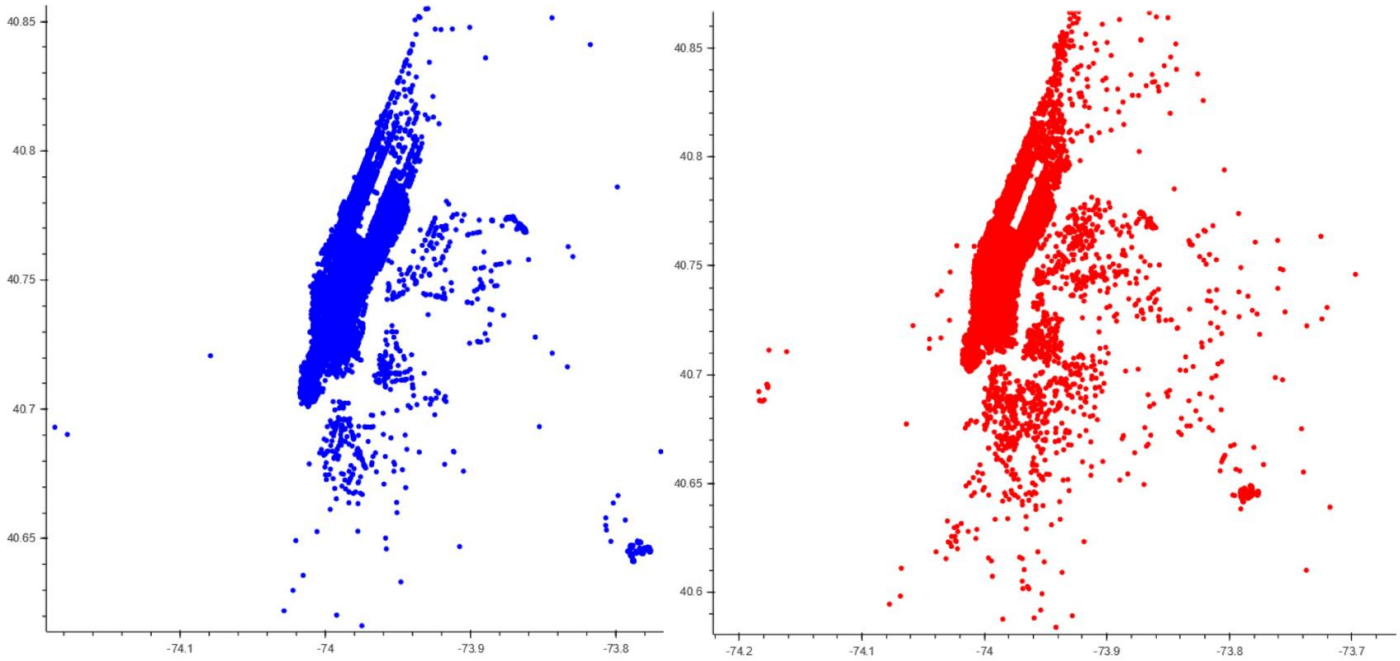

Figura 4 Plot de los puntos de recogida y llegada NYC Yellow

No se aprecian diferencias entre los distintos meses analizados. Hay diferencias entre las zonas de recogida y llegada entre los taxis verdes y amarillos, ya que los verdes no pueden recoger pasaje en las zonas de bajo Manhattan, central park hasta la zona de la bolsa (Wall Street). Se observa que la zona de recogida preferida son los barrios más cercanos a Manhattan, ya que dichos barrios constan con una fuerte presencia de recogidas de taxis.

Los taxis verdes no recogen pasaje en los aeropuertos.

1.3 Mejora la visualización anterior con un heat map.

Hemos observado que usando la librería datashader se realiza una buena visualización de mapa de calor, realizando una cartografía muy buena del terreno, con fuerte luz en las zonas de más influencia, es una buena librería a tener presente para datos con coordenadas de geolocalización.

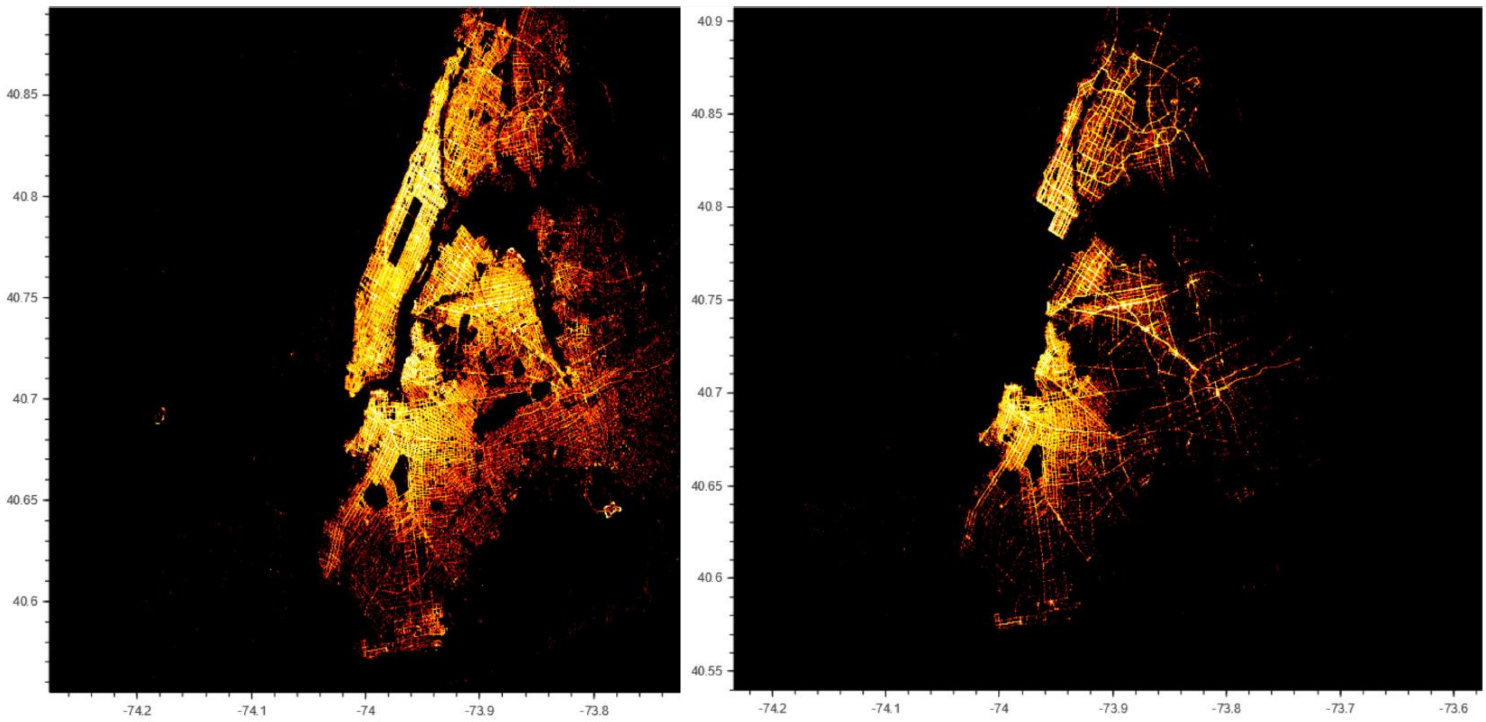

Figura 5 Heat map NYC Green

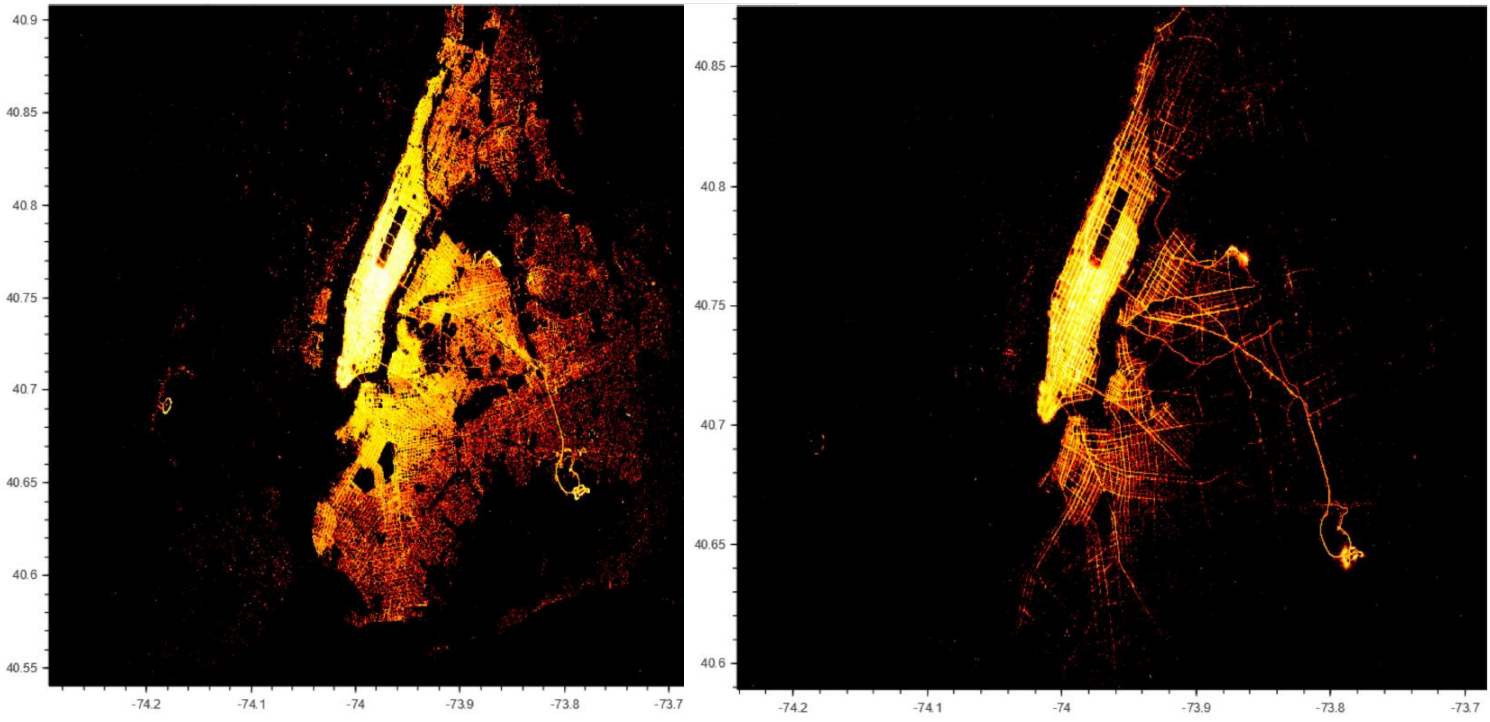

Figura 6 Heat map NYC Yellow

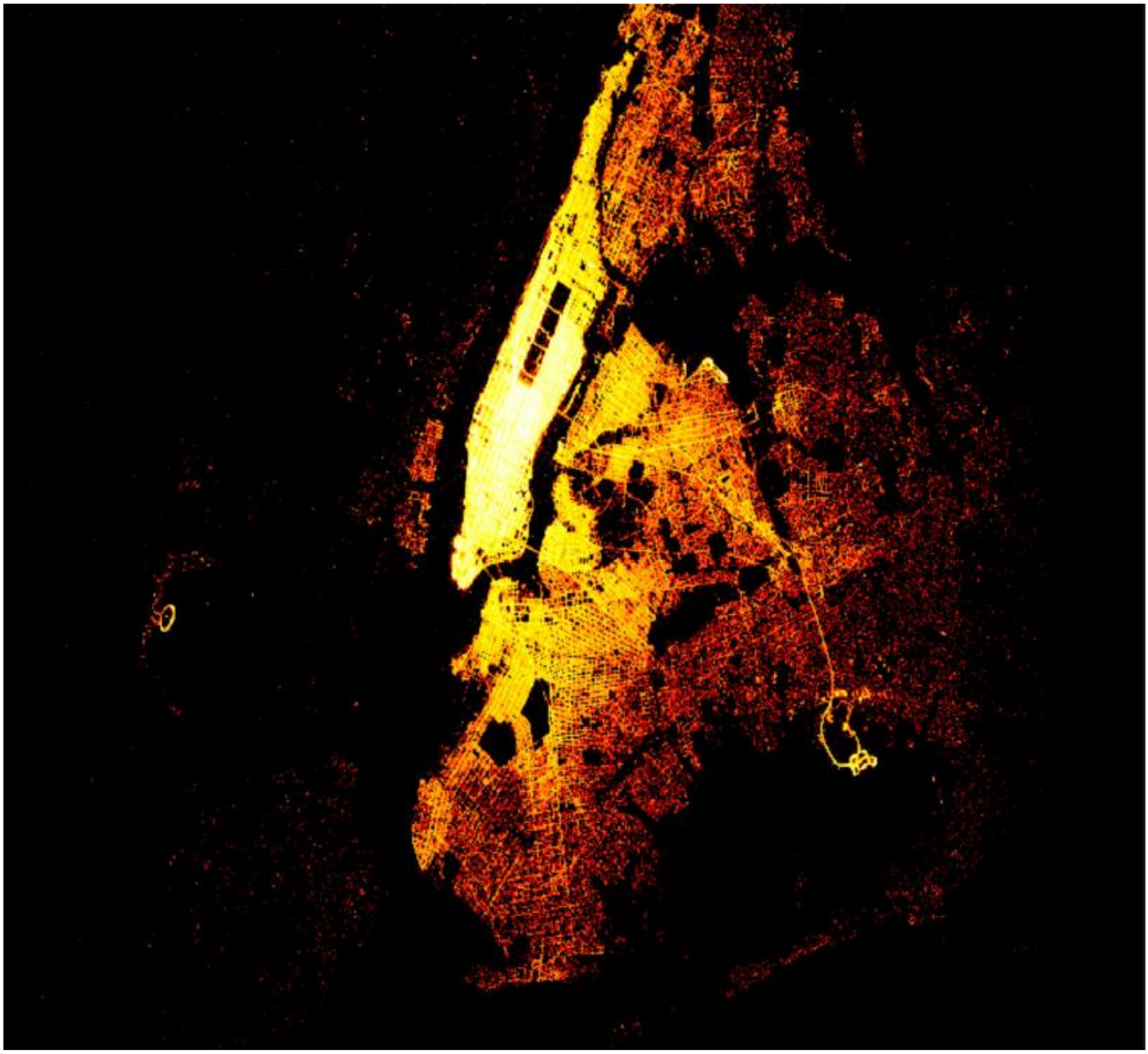

Figura 7 Heat Map NYC Yellow en detalle

2. Análisis Cualitativo

2.1 ¿Cuál es el trayecto en el que la relación precio/km es más alta? ¿Cuál es el trayecto en el que la relación tiempo/km es más alta? ¿Cuál es el trayecto en el que la relación precio/tiempo es más alta?

2.2 ¿Cuál es el trayecto en el que la relación precio/km es más baja? ¿Cuál es el trayecto en el que la relación tiempo/km es más baja? ¿Cuál es el trayecto en el que la relación precio/tiempo es más baja?

2.3 Muestra la evolución del tiempo medio de trayecto a lo largo del día. Muestra la evolución de la distancia media de trayecto a lo largo del día.

2.4 En la película "Jungla de Cristal 3: La venganza", Bruce Willis y Samuel L. Jackson deben ir en menos de 30 minutos desde la 72 con Broadway hasta la parada de metro de Wall Street (Broadway con Rector St) para evitar que estalle una bomba2. Calcula si esto es posible y en caso de que lo fuera, qué probabilidad de éxito tendrían (calculada como casos de éxito dividido por casos totales).

Los apartados correspondientes al punto 2, se resuelven en el código adjunto.

3. Análisis Predictivo

3.1 ¿Cuáles son las zonas donde es más probable coger un taxi en función de la hora del día?

Figura 8 Centroides clusters NYC Green

Figura 9 Clustering NYC Green

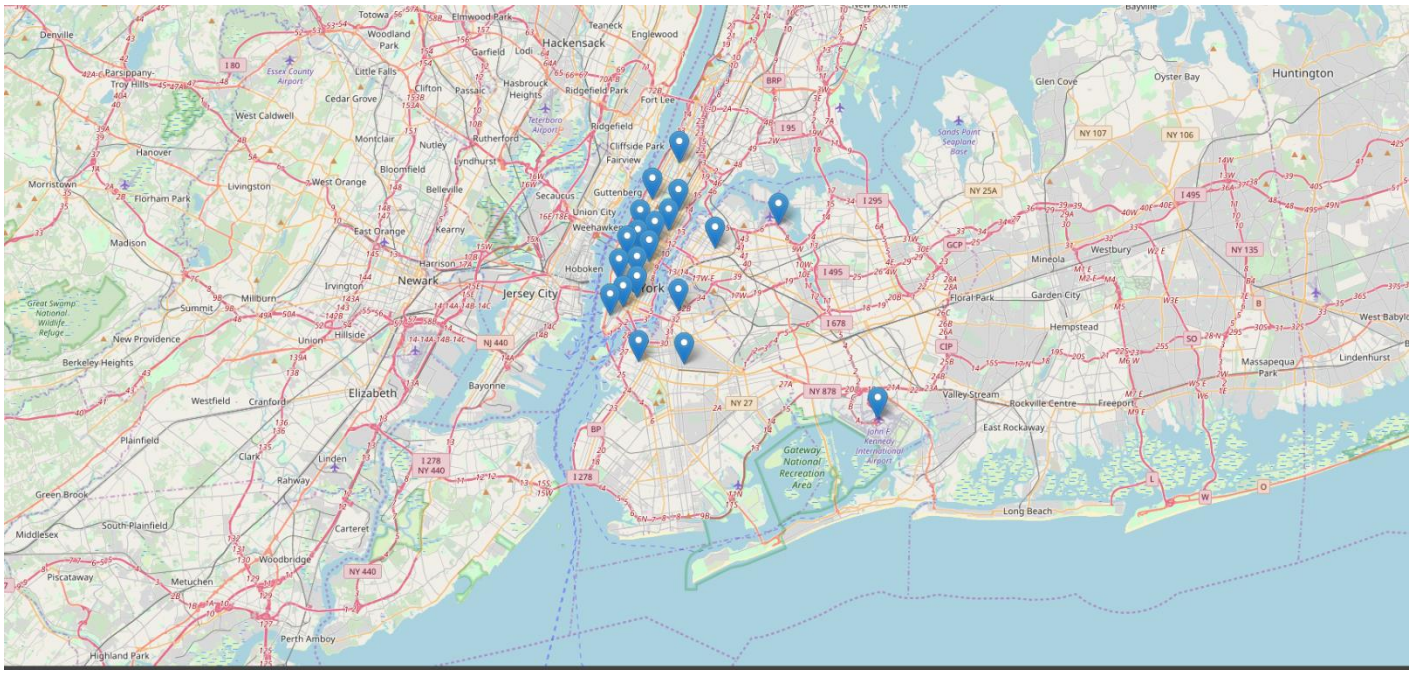

Figura 10 Centroides clusters NYC Yellow

Figura 11 Clustering NYC Yellow

Las zonas pertenecientes a los clústeres son las que tienen mayor probabilidad.

3.2 ¿Cuál es la mejor hora del día para ir al aeropuerto?

Aeropuerto	Mejor hora del día	Tiempo a aeropuerto	Distancia
JFK	14	123 min	1.83 millas
EWR	22	98 min	1.33 millas
LGA	21	53 min	1 milla

3.3 Diseña un modelo que, dada una hora, unas coordenadas origen, y unas coordenadas destino, predice la duración del trayecto y su coste. Muestra la relevancia de los atributos del dataset.

Apartado resuelto en el código anexo al trabajo. Se resuelve la predicción de total_amount sobre el dataset Green y la predicción de duración del trayecto sobre el dataset Yellow. No se realiza la predicción de ambas variables sobre ambos datasets, ya que ambas predicciones son reproducibles en ambos datasets.