

## NYC TAXI

Mateo Zapata & Ana María Montes & Nicolas Roza  
Septiembre 2021.

Universidad de los Andes.  
Departamento de Ingeniería de Sistemas.  
Ciencia de datos aplicada

## Tabla Contenido

Entender el problema .....	3
Ideación .....	4
Implicaciones éticas, riesgos legales o sesgos.....	5
Enfoque analítico.....	6
Entender los datos .....	6
Producto (Análisis descriptivo).....	6
Conclusiones y acciones sugeridas .....	8
Enlace repositorio .....	9
Referencias.....	9

Para el desarrollo del presente proyecto se propone abordar los datos de la agencia del gobierno de NYC Taxi & Limousine Commission respecto al servicio de taxis amarillos, con el fin generar una propuesta de valor, el trabajo se desarrollará mediante la metodología de ASUM-DM la cual para este primer sprint abarca hasta la etapa 5. Los roles para el trabajo se encuentran definidos de la siguiente manera:

**Líder del proyecto:** Ana María Montes.

**Líder de datos:** Nicolas Rozo Bedoya.

**Líder de modelos:** Mateo Zapata López.

## **Entender el problema**

Desde el año 2015 se observa una tendencia respecto a las tecnologías emergentes como las aplicaciones para servicio de transporte (Uber, Lyft, Juno, Via), con un aumento hasta del 300% en el año 2018 (Commission, What Makes a City Street Smart, 2019). Esta tendencia afecta a los prestadores de servicio tradicionales en especial en las grandes ciudades donde se han visto sujetos a competir en desventajas con estas grandes plataformas. Para el caso puntual de la ciudad de Nueva York en el año 2020, se estimaron un total de 250.000 conductores de las diferentes plataformas de servicio de los cuales el 95% son hombres entre los 25 y 54 años, el 91% de ellos son extranjeros, encabezados con un 13% de Dominica Republicana y 13% de Bangladesh (Commission, Annual Report , 2020).

La agencia del gobierno de NYC TLC (Taxi & Limousine Commission), otorgar licencias y regula la industria de taxis, vehículos de alquiler y aplicaciones, se encargan del establecimiento de las políticas de transporte público y son responsables directos de establecer las tarifas de taxis, supervisar la venta de los vehículos y su circulación. La agencia NYC le apuesta a la implementación de tecnologías y pilotos que permitan mejorar el servicio prestado y los beneficios obtenidos, dentro de sus principales intereses esta la recolección de datos, la ampliación y flexibilización de las características del parque automotor respecto a los tipos de combustible y modelos inclusivos para discapacitados. Los planes piloto se centran en los factores que afectan el comportamiento de conducción, las condiciones de trabajo, la tasa de colisión, la experiencia de los conductores, pasajeros y la relación de gastos y beneficios.

TLC tiene un monitoreo constante del servicio de taxis amarillo, por lo cual es consciente de su disminución en los últimos años debido a las tecnologías emergentes y la influencia del COVID-19. En la actualidad están interesados en evidenciar de manera cuantitativa la cantidad de personas y dinero reportados en los meses de diciembre en los últimos 4 años. TLC se pregunta que otros factores están involucrados en este comportamiento, que paso en los últimos 3 años antes

de la pandemia y en el año de pandemia y que estrategias se deben establecer para mejorar el servicio e incentivar el uso de taxis amarillos.

## Ideación

Con el fin de entender que piensa los involucrados del problema y cómo se comporta, se utilizó la herramienta de arquetipo para resumir lo recopilado en el entendimiento del problema como se presenta en la siguiente ilustración:

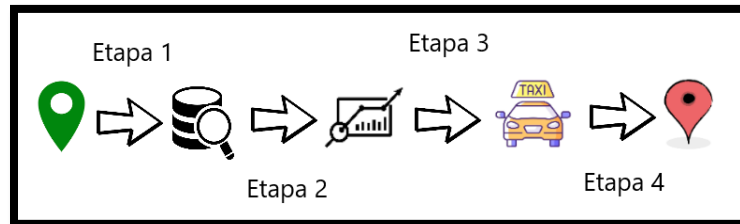


Ilustración 1 Arquetipos

Con los arquetipos definidos, se entiende las principales necesidades de los grupos de usuarios del tema a tratar. Por lo que se proponen las siguientes ideas para generar una estrategia que eviten la disminución del uso del servicio de taxis amarillos teniendo en cuenta el punto de vista del agente del gobierno.

1. Flexibilizar las políticas aplicadas a los taxistas.
2. Restablecer las tarifas que se cobran.
3. Erradicar otro tipo de servicios de transporte público o privado.
4. Viabilidad de uso por parte del usuario.
5. Generación de reportes y recomendaciones.
6. Horarios y lugares de circulación solo para taxis.

7. Incentivos a las personas que trabajan en el servicio de taxis.  
Para nuestro estudio se selecciona la idea de generación de reportes y recomendaciones para mejorar el servicio de taxis.



**Ilustración 2 Storyboard**

En una primera etapa se hace la recolección y búsqueda de fuentes de información. En la etapa 2 se realiza el análisis, limpieza y muestra de resultados en diferentes gráficos, finalmente en la etapa 3, sugerir a los *stakeholders*, maneras de mejorar el servicio de taxis amarillos.

### **Implicaciones éticas, riesgos legales o sesgos**

NYC Taxi & Limousine Commission (TLC) no se hace cargo de la recolección directa de la información solo de la recopilación y su disposición al público a través del portal de datos abiertos de la ciudad, no se hacen responsables por la veracidad de la información presentada. Para acceder a otro tipo de información interna, se realiza una solicitud directa avalada por la ley de información libre.

La TLC tiene como pilar garantizar el acceso al transporte de los usuarios discapacitados en especial quienes usan sillas de ruedas, por tal razón es controversial que se vean envueltos en problemas legales por negarse a permitir el uso de taxis ajustados para discapacitados por utilizar combustible alternativo (Chan, Sewell , 2005). Esto representa un sesgo sobre el uso de combustibles alternos y a su vez un conflicto, ya que si desean incentivar el uso del servicio de taxis están fallando al poner restricciones en las características del parque automotor. Adicionalmente la información provista por las aplicaciones como Uber, Lyft, Juno, Via no se monitorea de forma rigurosa aun cuando se sabe que el servicio con más inconvenientes, por su forma de operar y con la disminución de los salarios, los conductores se ven obligados a trabajar por más de 16 horas al día. Se considera que no realizar este análisis de forma rigurosa afecta el servicio de taxis ya que permite que otras plataformas tengan conductores con infracciones graves, licencias vencidas.

## Enfoque analítico

El análisis de datos del proyecto se encargará de examinar el conjunto de datos de taxis amarillos de NYC con el propósito de validar la disminución del uso del transporte de taxis amarillos en los últimos años debido a las tecnologías emergentes y la influencia del COVID-19.

De lo anterior, se define la siguiente hipótesis: Existe una disminución de los usuarios y los ingresos del servicio de taxis amarillos en los últimos años.

Para cumplir el objetivo se analizarán los factores correspondientes al volumen de pasajeros, ingresos totales, puntos más frecuentados de salida y zonas con más propinas. Con los resultados, se evaluarán diferentes rutas de acción para promover el uso del servicio de taxis amarillos y finalmente aumentar los ingresos que se generan por parte de este.

## Entender los datos

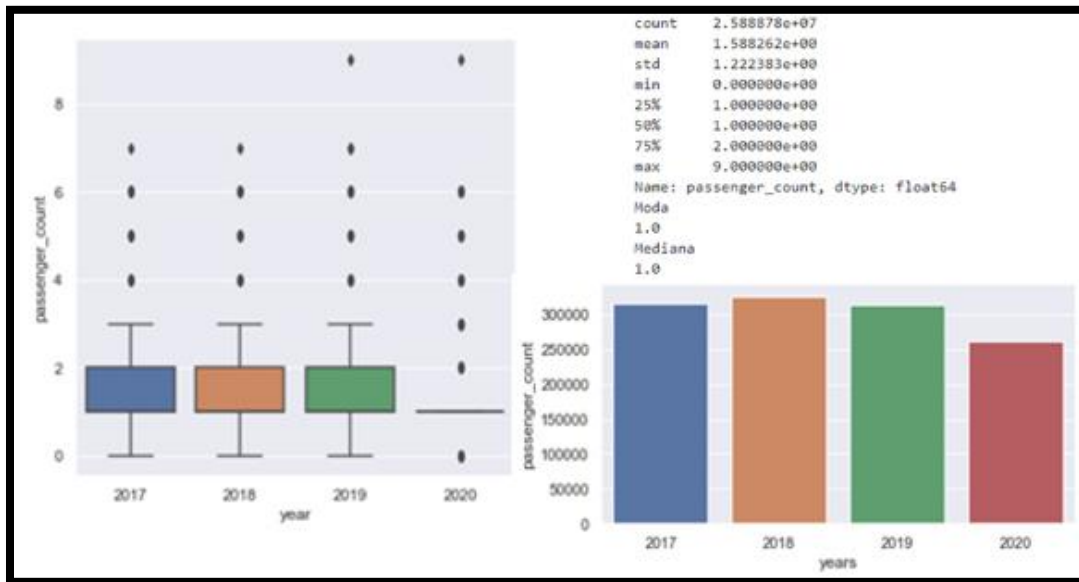
La agencia del gobierno de NYC Taxi & Limousine Commission cuenta con un repositorio de datos públicos en el cual se encuentran los registros de viaje de los servicios de transporte bajo su cobertura, los datos fueron proporcionados por proveedores de tecnología autorizados. El *dataset* a utilizar en el proyecto está ubicado en (Commission, New York City Taxi and Limousine Commission, 2021), allí se encuentra información de cada año y mes de los taxis amarillos que se movilizan en la ciudad de New York, se eligió utilizar los meses de diciembre de los años 2017, 2018, 2019 y 2020 para dar solución a la problemática planteada. El formato de los datos corresponde a un archivo csv, el cual contiene datos estructurados. El *data dictionary* de los archivos seleccionados se encuentra en (Data Dictionary Trip Records Yellow, 2021)[f](#), aquí se encuentra la información de las columnas que se encuentran en el *dataset*.

## Producto (Análisis descriptivo)

El “dataframe” principal se genera a partir de los cuatro archivos que contienen la información de los viajes que han realizado los taxis amarillos, se crea una nueva columna que contenga únicamente la fecha según la variable “tpep\_pickup\_datetime” que está en formato timestamp y que representa la fecha de salida. Luego, se aplica una eliminación de filas cuyo año no corresponden a la fecha del archivo. El motivo de este proceso es evitar los registros dañados. Por último, se incluye la descripción del nombre de la zona de salida y de llegada.

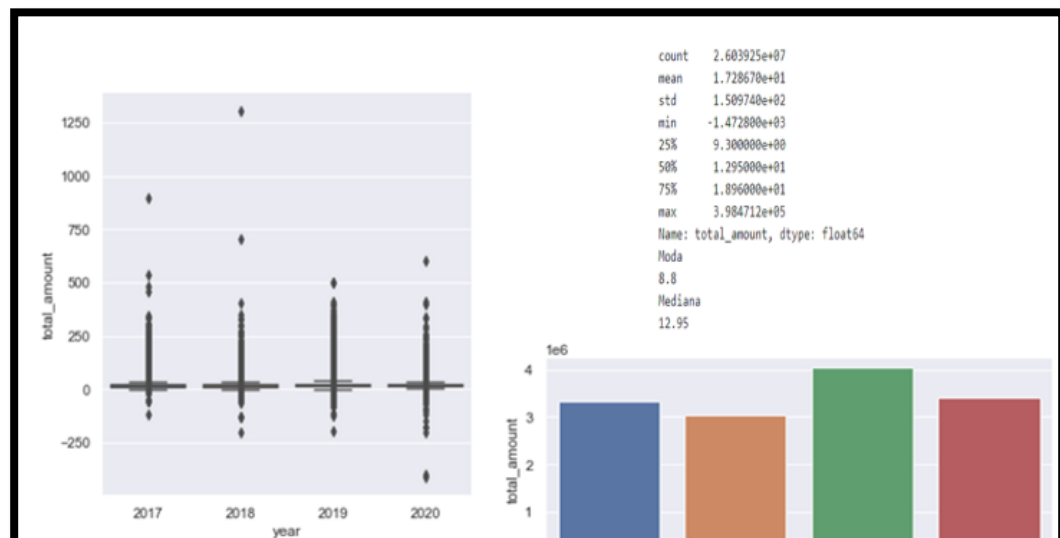
La ilustración 3 permite observar una disminución más marcada de la cantidad de pasajeros en el año 2020, la cual está ligada al virus Covid-19, asociadas a las medidas de confinamiento que afectaron la libre movilidad de las personas.

Calculando los porcentajes de pasajeros que se movizaron cada año obtenemos los siguientes resultados [2017 : 26%, 2018 : 27%, 2019 : 26%, 2020 : 21%].



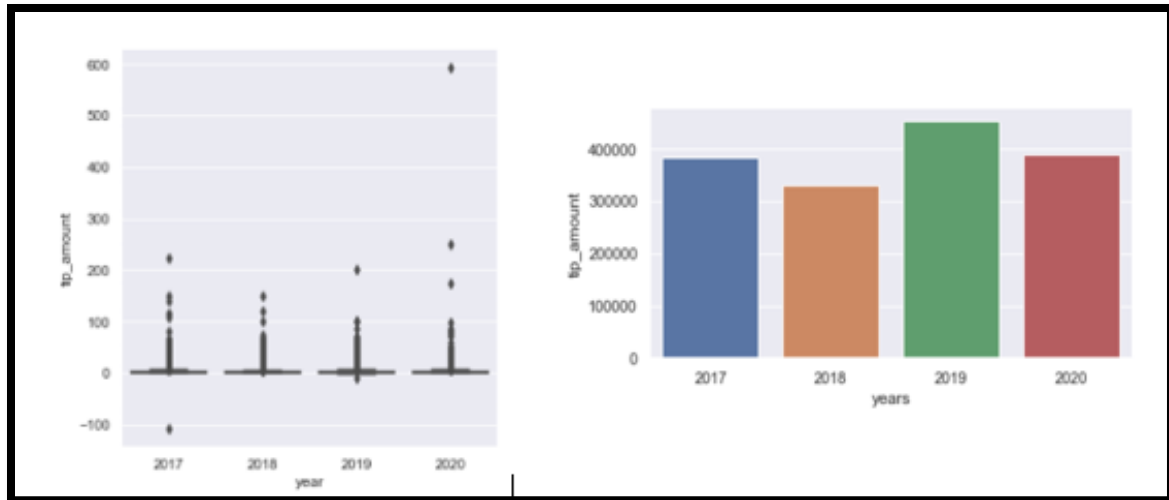
**Ilustración 3** Boxplot, medidas de tendencia, dispersión y diagrama de barras de la variable pasajeros

En la ilustración 4 se visualiza que el año con mayores ingresos fue el 2019, y también observamos una gran cantidad de *outliers*, que pueden ir ligados a un mal ingreso de registros, ya que un valor negativo no existe. Encontrando los porcentajes de la cantidad total de ingresos por año [2017 : 24%, 2018 : 22%, 2019 : 29%, 2020 : 25%]



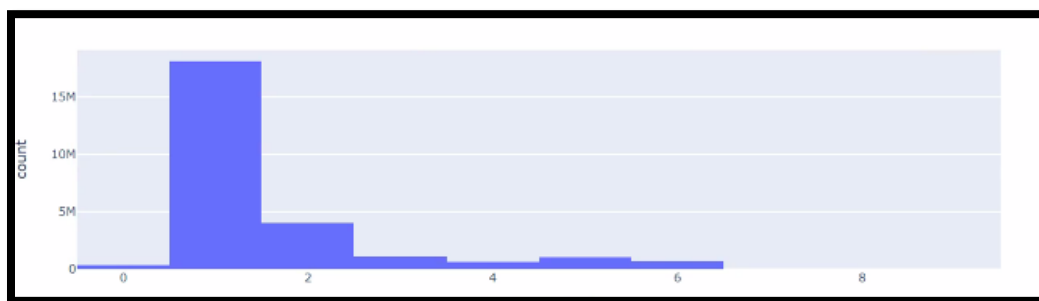
**Ilustración 4** Boxplot, medidas de tendencia, dispersión y diagrama de barras de la variable pasajeros

En la ilustración 5 podemos evidenciar que la tendencia del valor total va ligado al valor de las propinas, ya que son variables dependientes. Hallando los porcentajes del total de las propinas por año obtuvimos [2017 : 25%, 2018 : 21%, 2019 : 29%, 2020 : 25%].



**Ilustración 5 Boxplot y Diagrama de barras de la variable de propinas**

En la ilustración 6 el histograma muestra claramente la cantidad de pasajeros por viaje (Predomina 1 pasajero por viaje), por lo que se selecciona esta variable para realizar la prueba ANOVA detallada en el código.



**Ilustración 6 Histograma cantidad de pasajeros por viaje**

## Conclusiones y acciones sugeridas

Por medio del análisis descriptivo se corroboró la hipótesis planteada respecto a la disminución de los usuarios del servicio de taxis amarillos.



En el histograma realizado y las medidas de tendencia central para la cantidad de pasajeros predomina un pasajero por viaje, lo cual indica que una posible estrategia son viajes compartidos que permitan optimizar el rendimiento por servicio. Adicionalmente se construyó la hipótesis de la igualdad de medias para esta variable respecto a los años de estudio para la cual el análisis ANOVA rechaza la hipótesis nula.

Al explorar las otras variables de interés se encontró que la propina es proporcional al total de ingresos y no está ligada a la cantidad de pasajeros por viaje. Se tiene una alta cantidad de outliers que se deben analizar con más detalle en el siguiente sprint con el fin de identificar su procedencia.

### Enlace repositorio

El proyecto se trabajó en el repositorio de Github ([enlace](#)) que contiene el notebook. Adicionalmente se utilizó la funcionalidad de proyecto para hacer un seguimiento de cada una de las tareas realizadas, especificando sus características, responsable y etapa en la que se encuentra.

### Referencias

- Chan, Sewell . (4 de Noviembre de 2005). On the Street: Cleaner-Running Cabs". The New York Times. *The New York Times*. Obtenido de <https://www.nytimes.com/2005/11/04/nyregion/on-the-street-cleanerrunning-cabs.html>
- Commission, N. Y. (31 de Enero de 2019). What Makes a City Street Smart. Obtenido de <https://medium.com/@NYCTLC/what-makes-a-city-street-smart-23496d92f60d>
- Commission, N. Y. (2020). *Annual Report* . Annual Reppor , New York City Taxi and Limousine Commission, Office of Inclusion , New York . Obtenido de <https://www1.nyc.gov/assets/tlc/downloads/pdf/ooi-annual-report-2020.pdf>
- Commission, N. Y. (2021). *New York City Taxi and Limousine Commission*. Obtenido de <https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page>
- Data Dictionary Trip Records Yellow*. (2021). Obtenido de [https://www1.nyc.gov/assets/tlc/downloads/pdf/data\\_dictionary\\_trip\\_records\\_yellow.pdf](https://www1.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf)