# Ciencia de Datos Aplicada – Primera entrega del proyecto

Jose F. Corzo Manrique, Felipe A. Gutiérrez Naranjo, Alejandro Mantilla Redondo
Reporte técnico del primer *sprint* del proyecto
Universidad de los Andes, Bogotá, Colombia
{j.corzom, fa.gutierrez, a.mantillar}@uniandes.edu.co
Fecha de presentación: septiembre 20 de 2021

## Tabla de Contenido



## 1 Enlace al repositorio de solución

Esta sección contiene el enlace al repositorio específico donde se encuentra la solución de esta tarea: https://github.com/MantiMantilla/Ciencia-de-datos-aplicada-2021-20. En el repositorio se incluye un Jupyter Notebook que contiene más gráficos que contribuyen al análisis del proyecto.

## 2 Entendimiento del problema

En una era de competitividad fuera del gremio, una historia de violencia, robos y evasiones de pago, y un duro golpe por la pandemia global, el gremio de taxistas de la ciudad de Nueva York requiere de un nuevo distintivo competitivo con el cual asegurar mejores ingresos durante sus periodos de trabajo.

### 2.1 Datos del sector

El conjunto de datos que el grupo de trabajo decide utilizar para enfrentar el contexto de los conductores de taxi en la ciudad de Nueva York son los datos de registro de viajes de la TLC (*Taxi & Limousine Commission*) sobre sus taxis amarillos desde el año 2009. Los datos fueron recolectados por proveedores tecnológicos bajo el Programa de Mejora de Pasajeros de *Taxicab & Livery* (1).

### 2.2 Justificación

La profesión de un conductor en la ciudad de Nueva York ha contado históricamente, con serios obstáculos que han ido en aumento hasta el presente como casos de agresión y acoso por parte de pasajeros(2), la posibilidad que alguna persona no page al concluir con su trayecto y huya, forzando a los taxistas a reusarse a detenerse a pesar de que se considere ofensa laboral con posibles sanciones (4), y la competitividad recientemente inesperada con conductores de *Uber*, *Lyft* o similares con menor carga tributaria y falta de regulación clara (3). Estos factores hacen que la pérdida de ganancias y la desconfianza hacia sus clientes sean los aspectos más críticos para los conductores.

Como si lo anterior no fuera suficiente, la pandemia global del 2020 ha sido uno de los peores golpes para los taxistas de Nueva York, cortando la posible población de clientes activos debido a las leyes de cuarentena, y quienes desean movilizarse usualmente prefieren llamar a algún conductor por medio de una aplicación (5).

### 2.3 Problema por solucionar

Con base en la información anterior, el problema se designa como: **generar un plan de identificación de zonas y rutas que maximicen las ganancias de los taxistas, a partir del histórico de trayectos de taxi en la ciudad de Nueva York del TLC.**

### 2.4 Objetivos del proyecto

**Principal**

- Identificar los aspectos replicables de los trayectos más rentables en Nueva York, de forma que implementar consejos basados en ellos en un periodo de tres meses lleve a un incremento mínimo del 20% en las ganancias de los taxistas, para obtener una ventaja competitiva sobre otros conductores-por-contrato, con la implementación de un *dashboard* de apoyo.

**Específicos**

- Realizar un análisis conductivo al desarrollo de un *dashboard* verdaderamente útil para los taxistas de la ciudad de Nueva York.
- Crear modelos capaces de simular la implementación que puedan brindar información útil sobre qué decisiones tomar en pro de maximizar las ganancias.
- Desarrollar un prototipo de una herramienta tipo *dashboard*, posiblemente sujeta a cambio acorde a la información descubierta a lo largo del proyecto, que pueda dar solución a la situación.

### 2.5 Métricas de validación

Desarrollo de modelos predictivos que confirmen la correcta implementación de tal herramienta, junto al tipo de prueba apropiada para el análisis de datos aprendidos a lo largo de la clase.

## 3 Ideación

Una vez que se ha establecido el contexto de la situación y el horizonte de a dónde llegar, se comienza a concretar un plan de acción.

### 3.1 Arquetipos

Con base en la información encontrada, se considera que los arquetipos más apropiados son los conductores de taxi y los posibles pasajeros.

**Conductor de taxi:**

Hace:

- Espera por horas en lugares conocidos, como hoteles o aeropuertos, para intentar conseguir quien le pida una carrera.
- Ignora posibles carreras en la calle y realiza carreras con un alto nivel de estrés.

Piensa:

- Estos lugares son los únicos que conoce que tienen una alta probabilidad de resultar en carreras largas y, por ende, mayor ganancia.
- Es probable que estas personas se escapen sin pagar al final de la carrera, aunque nada más que el instinto de lo informa.
- No se tiene idea si este pasajero puede tornarse violento, lo que puede llevar a excluir ciertas zonas del todo, por miedo.

**Pasajero:**

Hace:

- Solicita a algún conductor en alguna de las varias aplicaciones para pedir que alguien lo lleve a algún lugar.
- Se para en la calle, solicitando a los taxis que pasan, pero ninguno se detiene.
- Cuando termina una carrera, no deja ninguna propina.

Piensa:

- Las aplicaciones son tan convenientes, que una leve diferencia de precio -mayor o menor- no importa tanto como la comodidad.
- No tiene idea por qué lo están ignorando en la calle; considerando que es obligatorio que los taxis paren cuando se les solicita, bien podría generar una queja con la TLC.
- La actitud del taxista, que no siga sus especificaciones o que el taxista se desvía -tenga un GPS activado o no- no dejan una buena impresión por la cual pagar extra.

### 3.2 Lluvia de ideas

Una aplicación con la cual los habitantes de Nueva York puedan solicitar taxis con la misma comodidad que solicitan a otros, mientras que los taxis siguen siendo los únicos que pueden recoger a pasajeros al azar en las calles.

Un *dashboard* con la capacidad de informar que zonas y que trayectos posibles, en efecto, tienen la mayor probabilidad de brindar buenas ganancias a lo largo de un día particular. **Esta es la idea seleccionada.**

Una nueva política, con base en consejos obtenidos tras un análisis de datos, que informen cómo hacer de los viajes más cómodos para el usuario, maximizando la oportunidad de que page una propina.

### 3.3 *Story board* del prototipo

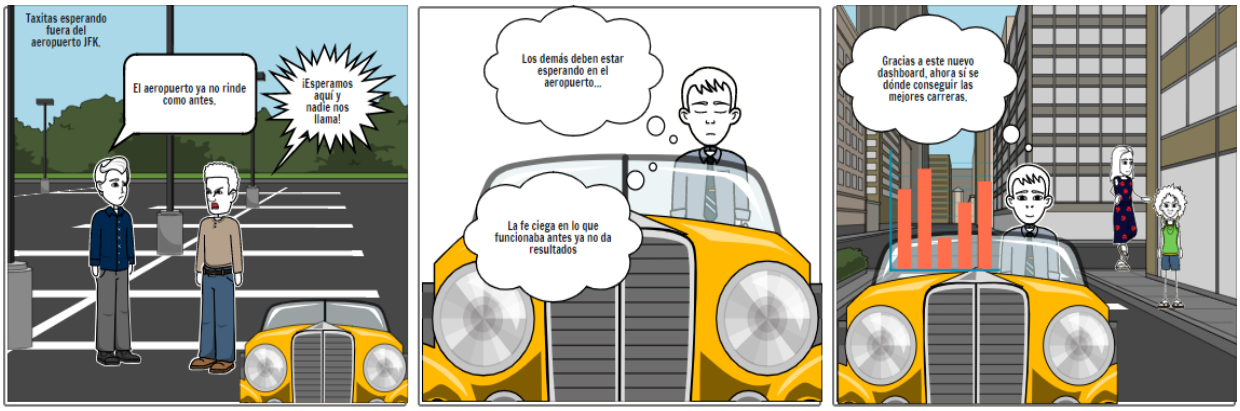


**Figura 1 - *Storyboard* de la solución planeada**

## 4 Consideraciones sociales

Con base en la información recolectada, se destacan las siguientes consideraciones. En cuanto a lo legal, el aspecto público de los datos los libera de líos de gobernanza o transparencia, pero es necesario que la herramienta no incite a rechazar servicios. Haciendo énfasis en la transparencia y la privacidad, se resalta que tanto los datos como el código del proyecto son abiertos. En cuanto a sesgos, se contempla la posibilidad que la información de criminalidad y características poblacionales de Nueva York generen sesgos contra ciertos sectores, como el Bronx, o etnias, como los afrodescendientes.

## 5 Enfoque analítico

Ahora se describe el enfoque que se realizará a lo largo del proyecto.

### 5.1 Preguntas a trabajar primer sprint

Se plantean las siguientes preguntas para el proyecto, en pro de adquirir información con la cual maximizar las ganancias de los conductores de taxi.

- Confirmar, por medio de las medias de dos años, si el monto recaudado por los taxistas ha cambiado.

### 5.2 Tipos de análisis predictivo a realizar

Se quiere verificar que las medias y desviaciones de distintos grupos de datos sean similares o no y a futuro modelos de regresión, clasificación y su uso para predecir.

## 6 Entendimiento de los datos

Como se menciona en la sección 2.1, el conjunto de datos proviene del sitio público de la TLC. Los datos son de carácter semiestructurado, almacenados en archivos `.csv` representantes de cada mes de un año dado. A pesar del cambio de ciertos nombres y tipos de datos a lo largo de los años, las columnas del conjunto para los taxis son:

- **VendorID**: código indicativo del proveedor TPEP (Trusted Product Evaluation Program).
- **tpep_pickup_datetime**: la fecha exacta en que empezó el trayecto.
- **tpep_dropoff_datetime**: la fecha exacta en que terminó el trayecto.
- **Passenger_count**: el número de pasajeros en el trayecto.

- **Trip_distance**: la distancia que tomó el trayecto.
- **PULocationID**: identificador de la zona en la que empezó el trayecto.
- **DOLocationID**: identificador de la zona en la que terminó el trayecto.
- **RateCodeID**: código de recargas al costo del trayecto por su tipo.
- **Store_and_fwd_flag**: indicador del estado de conexión entre el taxi y el servidor.
- **Payment_type**: código del tipo con el que se pagó el trayecto.
- **Fare_amount**: el costo base del trayecto.
- **Extra**: costos adicionales misceláneos.
- **MTA_tax**: un impuesto que puede aplicar al trayecto.
- **Improvement_surcharge**: otro modificador al costo del trayecto.
- **Tip_amount**: la cantidad de propina recibida por el trayecto.
- **Tolls_amount**: total pagado en peajes por el trayecto.
- **Total_amount**: el total que se cobró por el trayecto.

La calidad de los datos, en general, es bastante adecuada; se cuenta con una cantidad adecuada de columnas que permiten plantear varias hipótesis posibles, sin mencionar el volumen masivo de filas con las que se cuenta.

# 7 Tratamiento de datos

La sección muestra los supuestos bajo los cuales se planea trabajar con los datos.

## 7.1 Proceso de limpieza

En pro de limitar el alcance del proyecto, se plantea trabajar con los datos de 2017 en adelante, ya que son estos años los que cuentan con un identificador claro de las zonas en las que ocurren los trayectos, en vez de datos georreferenciados. Adicionalmente, se realiza un proceso de selección de columnas con las cuales trabajar: aquellas relacionadas con fechas se reemplazan con las columnas *year*, *month*, *week_day*, *month_day*, *hour* y *season* con información extraída/calculada de las primeras. En el mismo orden de ideas, las columnas *VendorID*, *RateCodeID*, Store_and_fwd_flag, *Extra*, *MTA_tax*, *Improvement_surcharge* y *congestion_surcharge* (una columna específica del 2017) son eliminadas.

## 7.2 Tecnología relevante

Como parte del tratamiento de semejante volumen de datos (325.000.000+ filas) y como posible facilitador a futuro, se opta por utilizar una solución más robusta por servicios de terceros. Esta arquitectura radica en *Dask*, un marco de trabajo para computo distribuido de ciencia de datos, que integra varias librerías de *Python* de forma nativa -como *Pandas*, *NumPy* y *Scikit-learn*- y maneja sus colecciones de datos por medio de un *dataframe* distribuido en cada nodo (los cuales asignan la cantidad de recursos necesarios sin desperdicio por medio de *elastic computing*) gracias a una unión que realiza sobre múltiples *dataframes* llevados casi a su tamaño límite, sobre un clúster de procesamiento de *big data* de Amazon EMR (parte de los *Amazon Web Services*), donde se destaca cómo **no** se necesita de almacenamiento gracias a que el conjunto de datos seleccionado es uno de aquellos que AWS provee (6).

## 7.3 Análisis descriptivo

La siguiente tabla tiene las estadísticas descriptivas de las variables más relevantes del ejercicio.

| | passenger_count | trip_distance | total_amount | trip_dur_secs |
|---|---|---|---|---|
| **count** | 3.242907e+08 | 3.253469e+08 | 3.253469e+08 | 3.253469e+08 |
| **mean** | 1.587871e+00 | 2.994856e+00 | 1.724392e+01 | 1.036081e+03 |
| **std** | 1.232861e+00 | 9.023515e+01 | 1.934956e+02 | 3.909878e+03 |
| **min** | 0.000000e+00 | -3.726453e+04 | -1.871800e+03 | 0.000000e+00 |
| **25%** | 1.000000e+00 | 1.130000e+00 | 1.130000e+01 | 5.150000e+02 |
| **50%** | 1.000000e+00 | 2.240000e+00 | 1.580000e+01 | 8.680000e+02 |
| **75%** | 2.000000e+00 | 9.540000e+00 | 4.300000e+01 | 2.220000e+03 |
| **max** | 1.920000e+02 | 3.509149e+05 | 1.084772e+06 | 8.639900e+04 |

Figura 2 - Tabla descriptiva

Entre los gráficos más relevantes realizando en Plotly se encuentran resultados llamativos para el planteamiento de hipótesis:

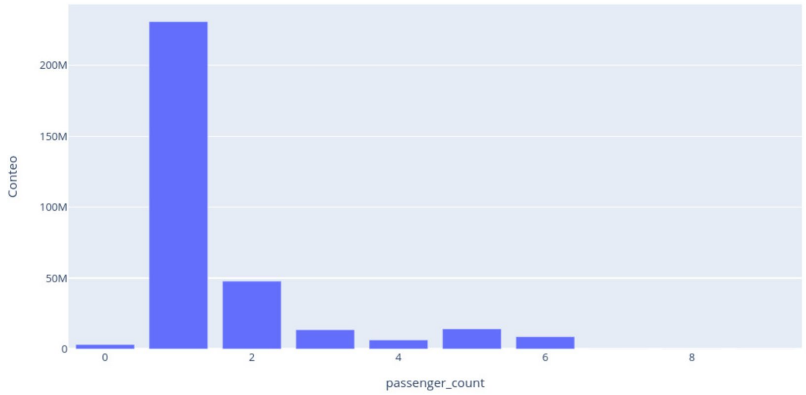


Figura 3 - Histograma número de pasajeros por viaje

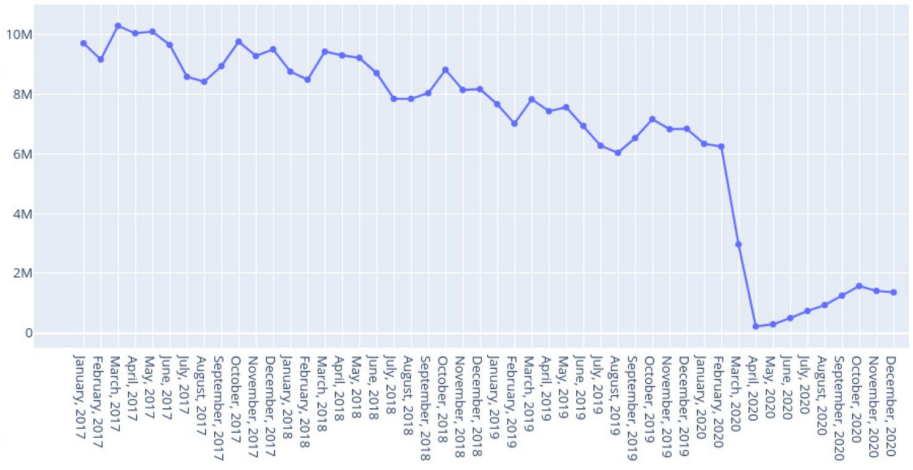


Figura 4 - Total de viajes por mes

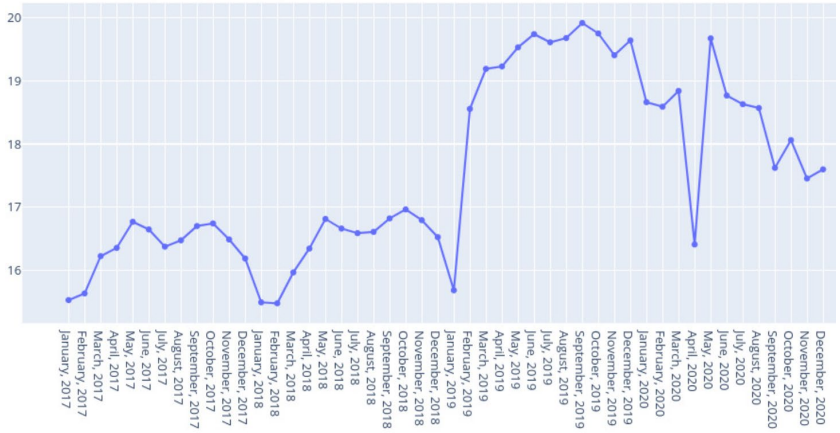


Figura 5 - Promedio de tarifa por mes

### 7.4 Validación de hipótesis

Vamos a analizar la diferencia entre la media de la tarifa entre los años 2017 y 2019.

- $H_0 : \mu_{2017} = \mu_{2019}$
- $H_1 : \mu_{2017} \neq \mu_{2019}$

Valor de t calculado: -111.85

Valor p: 0.0

Por lo tanto, se rechaza la hipótesis nula.

## 8 Cierre de *sprint*

Después de lo que se consigue con este *sprint*, esta sección representa las consideraciones de cierre de esta entrega.

### 8.1 Primeras conclusiones

- La inclusión de un clúster para análisis de *big data* es beneficioso pero los tiempos de ejecución pueden subestimarse y el costo por el número de *workers* puede ser más del deseado.
- El realizar los análisis con todo el conjunto de datos, en lugar de realizar un muestreo, permite realizar los cálculos con la mayor precisión posible, pero el procesamiento de estos genera un gran obstáculo en términos de infraestructura y tiempo.
- Se nota que el total de pasajeros transportados por taxis amarillos entre el 2017 y el 2020 es de 324 millones, mientras que el total recaudado fue de 325 millones de dólares.
- En promedio los viajes tienen una duración de 17 minutos y el 50% de estos se realizan entre 9 y 37 minutos.
- Con un nivel de confianza del 95% se concluye que la media de los años 2017 y 2019 no son iguales confirmando que el monto recaudado por los taxistas ha cambiado.

### 8.2 Acciones sugeridas para el siguiente sprint.

- Realizar una segunda ronda de limpieza de datos de valores atípicos encontrados en la tabla descriptiva.

- Realizar el procesamiento en el clúster con más días de antelación para procesos más complejos.
- Considerar realizar un muestreo del conjunto de datos estadísticamente significativo, pero computacionalmente viable.
- Desarrollo de modelos predictivos que confirmen la correcta implementación del *dashboard* junto al tipo de prueba apropiada para el análisis de datos aprendidos a lo largo de la clase.

# 9 Contribuciones

Se trabajó con una metodología de trabajo en vivo, donde cada miembro lideró una fase diferente, pero todos participaron: Felipe Gutiérrez dirigió la búsqueda de información, la ideación de soluciones y la creación del story board, Jose Corzo dirigió la selección de datos y el proceso de limpieza, y Alejandro Mantilla dirigió la configuración del clúster utilizado y la implementación de la lógica para los análisis.

# 10 Bibliografía


1. NYC Taxi & Limousine Commission. *TLC Trip Record Data.* [En línea] [Citado el: 20 de septiembre de 2021.] https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

2. Kloberdanz, Kristin. Health Day. *Taxi Drivers: Years of Living Dangerously.* [En línea] 31 de diciembre de 2020. [Citado el: 20 de septiembre de 2021.] https://consumer.healthday.com/encyclopedia/work-and-health-41/occupational-health-news-507/taxi-drivers-years-of-living-dangerously-646377.html.

3. Crudele, John. New York Post. *The challenges of driving a yellow cab in the age of Uber.* [En línea] 17 de septiembre de 2017. [Citado el: 20 de septiembre de 2021.] https://nypost.com/2017/09/18/the-challenges-of-driving-a-taxi-in-the-age-of-uber/.

4. —. New York Post. *Most cab drivers break this law for their own good.* [En línea] 21 de septiembre de 2017. [Citado el: 20 de septiembre de 2021.] https://nypost.com/2017/09/21/most-cab-drivers-break-this-law-for-their-own-good/.

5. Khan, Tasmiha. Streetblog Chicago. *Taxi drivers are struggling during the pandemic. Their rights need to be addressed.* [En línea] 24 de marzo de 2020. [Citado el: 20 de septiembre de 2021.] https://chi.streetsblog.org/2020/03/24/taxi-drivers-are-struggling-during-the-pandemic-their-rights-need-to-be-addressed/.

6. Vittal, Ram y Muppala, Sireesha. Amazon Web Services. *Machine learning on distributed Dask using Amazon SageMaker and AWS Fargate.* [En línea] 18 de febrero de 2021. [Citado el: 21 de septiembre de 2021.] https://aws.amazon.com/blogs/machine-learning/machine-learning-on-distributed-dask-using-amazon-sagemaker-and-aws-fargate/.