

Documento de Estrategia PACE

Proyecto Automatidata — TLC

EDA y Visualización de Datos

PACE: Etapa de Planificación

Las columnas más relevantes para el análisis son trip_distance, fare_amount, tip_amount, total_amount, tpep_pickup_datetime, tpep_dropoff_datetime, PULocationID, DOLocationID, VendorID, payment_type y passenger_count. Estas variables permiten analizar distancia, tiempo, montos del viaje, propinas, patrones temporales y diferencias por ubicación o proveedor.

La distancia del viaje está medida en millas y los montos monetarios en dólares. Las variables temporales corresponden a fecha y hora, mientras que passenger_count representa el número de pasajeros por viaje.

Como hipótesis iniciales, se asume que la mayoría de los viajes son cortos y de bajo costo, con distribuciones sesgadas a la derecha y presencia de valores atípicos. Asimismo, se espera identificar patrones temporales por mes y día, así como diferencias por proveedor y ubicación.

No se identificaron valores faltantes en el conjunto de datos. Todas las columnas presentan el mismo número de registros no nulos.

Las columnas numéricas presentan formatos consistentes. Las variables de fecha y hora requieren conversión a formato datetime para su correcto análisis.

Las prácticas de EDA necesarias incluyen revisión de estructura del dataset, análisis de distribuciones, detección de valores atípicos, análisis temporal y exploración de relaciones entre variables clave.

PACE: Etapa de Análisis

El EDA se realiza validando tipos de datos y rangos, explorando distribuciones y valores atípicos, analizando relaciones entre distancia y monto, e identificando patrones temporales y espaciales relevantes.

No fue necesario integrar fuentes de datos adicionales. Se aplicaron procesos de filtrado de valores no válidos, ordenación cronológica y creación de variables derivadas como mes y día.

Las visualizaciones se diseñaron para una audiencia no técnica, priorizando claridad y simplicidad mediante histogramas, diagramas de caja, gráficos de barras, series temporales y mapas.

PACE: Etapa de Construcción

Para cumplir los objetivos del proyecto se desarrollaron visualizaciones de EDA, incluyendo histogramas, diagramas de caja, gráficos de barras, gráficos temporales y un diagrama de dispersión. No se implementaron algoritmos de aprendizaje automático.

La construcción de visualizaciones incluyó limpieza y validación de datos, transformación de variables temporales, agregaciones por tiempo y categoría, y selección de escalas y etiquetas claras.

Las variables más relevantes para las visualizaciones fueron trip_distance, total_amount, tip_amount, month, day, VendorID, PULocationID y DOLocationID.

PACE: Etapa de Ejecución

El análisis mostró que la mayoría de los viajes son cortos y de bajo costo, con distribuciones sesgadas y presencia de valores atípicos. Se identificaron patrones temporales y diferencias por proveedor y ubicación.

A partir de estos hallazgos, se recomienda fortalecer controles de calidad de datos, utilizar los patrones identificados para mejorar estimaciones de tarifas y enfocar análisis operativos en ubicaciones con mayor volumen o mayor distancia promedio.

El proyecto puede ampliarse analizando el impacto de la hora del día o del método de pago en los montos totales, así como diferencias más detalladas entre zonas específicas.

Los resultados pueden comunicarse mediante notebooks técnicos para audiencias analíticas y dashboards interactivos en Tableau y Power BI para stakeholders no técnicos.