# Proyecto de curso

¡Enhorabuena! Has completado la sección de Recopilación y almacenamiento de datos. Es hora de aplicar el conocimiento y las habilidades que has adquirido en un proyecto: un estudio de caso analítico real que completarás por tu cuenta.
Cuando finalices el proyecto, envía tu trabajo al revisor de proyecto para su evaluación. Te dará su opinión en 24 horas. Utiliza los comentarios para realizar cambios y luego envía la nueva versión al revisor del proyecto.
Puede que recibas aún más feedback en la nueva versión. Esto es totalmente normal. Es común pasar por varios ciclos de comentarios y revisiones.
Tu proyecto se considerará completado una vez que el revisor del proyecto lo apruebe.

**Descripción del proyecto**

Estás trabajando como analista para Zuber, una nueva empresa de viajes compartidos que se está lanzando en Chicago. Tu tarea es encontrar patrones en la información disponible. Quieres comprender las preferencias de los pasajeros y el impacto de los factores externos en los viajes.

Al trabajar con una base de datos, analizarás los datos de los competidores y probarás una hipótesis sobre el impacto del clima en la frecuencia de los viajes.

## Descripción de los datos

Una base de datos con información sobre viajes en taxi en Chicago:

tabla `neighborhoods`: datos sobre los barrios de la ciudad

- `name`: nombre del barrio
- `neighborhood_id`: código del barrio

tabla `cabs`: datos sobre los taxis

- `cab_id`: código del vehículo
- `vehicle_id`: ID técnico del vehículo
- `company_name`: la empresa propietaria del vehículo

tabla `trips`: datos sobre los viajes

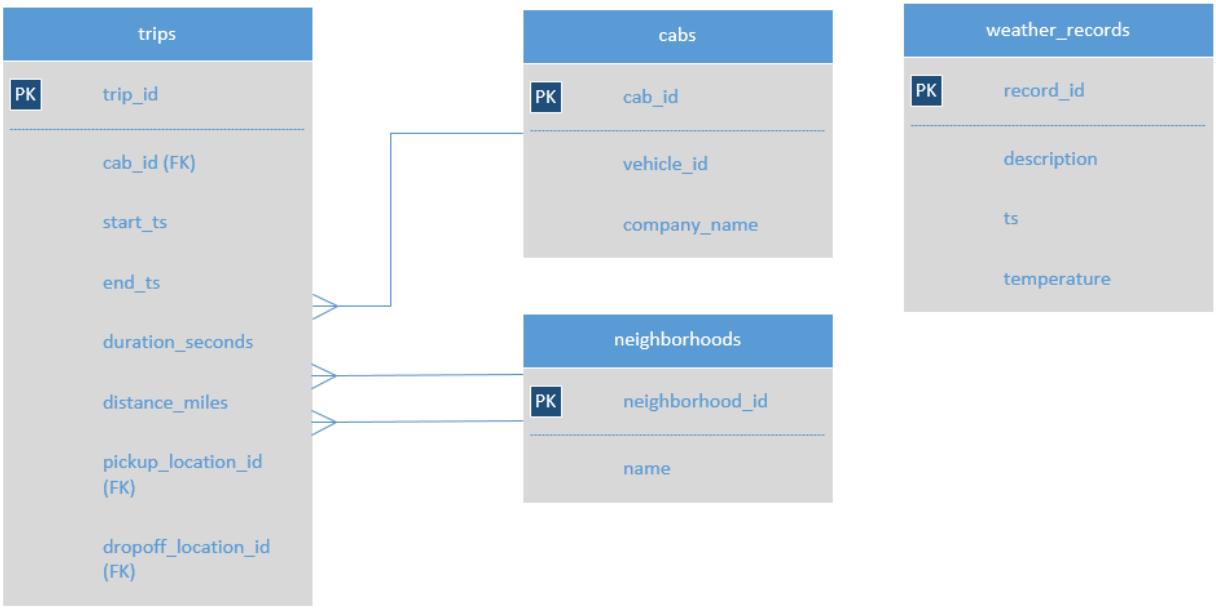
- `trip_id`: código del viaje

- `cab_id`: código del vehículo que opera el viaje
- `start_ts`: fecha y hora de inicio del viaje (tiempo redondeado a la hora)
- `end_ts`: fecha y hora de finalización del viaje (tiempo redondeado a la hora)
- `duration_seconds`: duración del viaje en segundos
- `distance_miles`: distancia del viaje en millas
- `pickup_location_id`: código del barrio de recogida
- `dropoff_location_id`: código del barrio de finalización del recorrido

tabla `weather_records`: datos sobre el clima

- `record_id`: código del registro meteorológico
- `ts`: fecha y hora del registro (tiempo redondeado a la hora)
- `temperature`: temperatura cuando se tomó el registro
- `description`: breve descripción de las condiciones meteorológicas, por ejemplo, "lluvia ligera" o "nubes dispersas"

## Esquema de tabla



Nota: no existe una conexión directa entre las tablas `trips` y `weather_records` en la base de datos. Pero aún puedes usar JOIN y vincularlas usando la hora en la que comenzó el viaje (`trips.start_ts`) y la hora en la que se tomó el registro meteorológico (`weather_records.ts`).

## Instrucciones para completar el proyecto

**Paso 1. Escribe un código para analizar los datos sobre el clima en Chicago en noviembre de 2017 desde el sitio web:**

http://slava-public-access.s3-website-eu-west-1.amazonaws.com

> https://s3-us-west-2.amazonaws.com/secure.notion-static.com/d752cb24-f5ed-4b1c-b6a4-570fbc905b7c/chicago_weather_2017.html

**Paso 2. Análisis exploratorio de datos**

1. Encuentra el número de viajes en taxi para cada empresa de taxis del 15 al 16 de noviembre de 2017. Nombra el campo resultante `trips_amount` y muéstralo junto con el campo `company_name`. Ordena los resultados por el campo `trips_amount` en orden descendente.
2. Encuentra la cantidad de viajes para cada empresa de taxis cuyo nombre contenga las palabras "Yellow" o "Blue" del 1 al 7 de noviembre de 2017. Nombra la variable resultante `trips_amount`. Agrupa los resultados por el campo `company_name`.
3. En noviembre de 2017 las empresas de taxis más populares fueron Flash Cab y Taxi Affiliation Services. Encuentra el número de viajes de estas dos empresas y nombra la variable resultante `trips_amount`. Junta los viajes de todas las demás empresas en el grupo "Other". Agrupa los datos por nombres de empresas de taxis. Nombra el campo con nombres de empresas de taxis `company`. Ordena el resultado en orden descendente por `trips_amount`.

**Paso 3. Prueba la hipótesis de que la duración de los viajes desde el Loop hasta el Aeropuerto Internacional O'Hare cambia los sábados lluviosos.**

4. Recupera los identificadores de los barrios de O'Hare y Loop de la tabla `neighborhoods`.

5. Para cada hora recupera los registros de condiciones meteorológicas de la tabla `weather_records`. Usando el operador CASE, divide todas las horas en dos grupos: "Bad" si el campo `description` contiene las palabras "rain" o "storm" y "Good" para los demás. Nombra el campo resultante `weather_conditions`. La tabla final debe incluir dos campos: fecha y hora (*ts*) y `weather_conditions`.

6. Recupera de la tabla `trips` todos los viajes que comenzaron en el Loop (`neighborhood_id`: 50) y finalizaron en O'Hare (`neighborhood_id`: 63) un sábado. Obtén las condiciones climáticas para cada viaje. Utiliza el método que aplicaste en la tarea anterior. Recupera también la duración de cada viaje. Ignora los viajes para los que no hay datos disponibles sobre las condiciones climáticas.

**Paso 4. Análisis exploratorio de datos (Python)**

Además de los datos que recuperaste en las tareas anteriores te han dado un segundo archivo. Ahora tienes estos dos CSV:

`project_sql_result_01.csv`. Contiene los siguientes datos:

- `company_name`: nombre de la empresa de taxis
- `trips_amount`: el número de viajes de cada compañía de taxis el 15 y 16 de noviembre de 2017.

`project_sql_result_04.csv.` Contiene los siguientes datos:

- `dropoff_location_name`: barrios de Chicago donde finalizaron los viajes
- `average_trips`: el promedio de viajes que terminaron en cada barrio en noviembre de 2017.

Para estos dos datasets ahora necesitas:

- importar los archivos
- estudiar los datos que contienen
- asegurarte de que los tipos de datos sean correctos
- identificar los 10 principales barrios en términos de finalización del recorrido

- hacer gráficos: empresas de taxis y número de viajes, los 10 barrios principales por número de finalizaciones
- sacar conclusiones basadas en cada gráfico y explicar los resultados

**Paso 5. Prueba de hipótesis (Python)**

`project_sql_result_07.csv`: el resultado de la última consulta. Contiene datos sobre viajes desde el Loop hasta el Aeropuerto Internacional O'Hare. Recuerda, estos son los valores de campo de la tabla:

- `start_ts`: fecha y hora de recogida
- `weather_conditions`: condiciones climáticas en el momento en el que comenzó el viaje
- `duration_seconds`: duración del viaje en segundos

Prueba la hipótesis:
"La duración promedio de los viajes desde el Loop hasta el Aeropuerto Internacional O'Hare cambia los sábados lluviosos".

Establece el valor del nivel de significación (alfa) por tu cuenta.

Explica:

- cómo planteaste las hipótesis nula y alternativa
- qué criterio usaste para probar las hipótesis y por qué

# ¿Cómo será evaluado mi proyecto?

Estos son los criterios de evaluación del proyecto. Léelos atentamente antes de empezar a trabajar.

Esto es lo que buscará el revisor del proyecto al evaluar tu proyecto:

- cómo recuperas los datos del sitio web
- cómo creas slices de datos
- cómo agrupas los datos
- si utilizas los métodos correctos para unir tablas
- cómo formulas las hipótesis

- qué criterios utilizas para probar las hipótesis y por qué
- a qué conclusiones llegas
- si dejas comentarios en cada paso

Las hojas informativas y resúmenes de las lecciones anteriores tienen todo lo que necesitas para completar el proyecto.

¡Buena suerte!