# Taxi & Weather Insights: Understanding Taxi Demand Trends and Weather Impact

November 26, 2024

## Question

How do weather conditions impact the frequency and duration of taxi trips in New York City?

## Data Sources

### New York Taxi Trip Data (2019)

- **Source:** Kaggle Dataset
- **Description:** This dataset contains detailed information on yellow taxi rides in New York City during 2019, including trip distances, passenger counts, and pickup and drop-off time stamps.
- **Structure:** Monthly raw CSV files, each including millions of rows.
- **License:** Open Data Commons Public Domain Dedication and License (PDDL).
- **Reason for Use:** This data provides detailed insights into transportation patterns in New York City.

### Weather Data

- **Source:** Visual Crossing Weather API
- **Description:** Temperature, wind speed, humidity, and visibility belong to the daily environmental data for New York City provided in this dataset.
- **Structure:** CSV file providing data on the weather for every single day in 2019.
- **License:** Visual Crossing provides free climate data access for academic and research uses. Proper credit must be given when using Visual Crossing's data in any shared or published work.
- **Reason for Use:** Weather is hypothesized to affect trip patterns and taxi demand. This dataset enables an investigation into the relationship between weather metrics and transportation trends.

- **Compliance:** The project ensures proper credit is given to comply with Visual Crossing's licensing terms:

  "Visual Crossing provided the weather data. https://www.visualcrossing.com."

# Data Pipeline

## Technology Stack

The following Python tools were used to implement the pipeline:

- **Data Retrieval:** Kaggle API and Visual Crossing API
- **Data Processing:** SQLite for interim storage, Pandas for manipulating data
- **Error Handling:** Exception management for processing and file download errors

## Data Pipeline Diagram

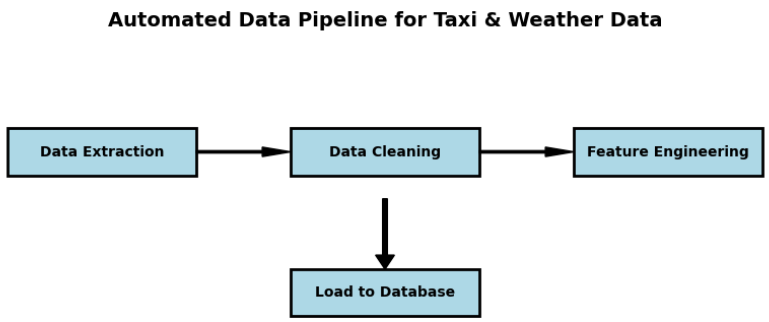


Figure 1: Automated Data Pipeline for Taxi & Weather Data

## Pipeline Description

- **Data Extraction:** Taxi data was downloaded from Kaggle and weather data from Visual Crossing.
- **Data Editing:** Taxi data was processed in chunks to handle memory limits. Relevant columns (e.g., number of passengers, timestamps) were filtered, and trip duration was calculated. Weather data was filtered to retain temperature, humidity, wind speed, and visibility.
- **Data Transformation:** Both datasets were aggregated by date and saved as separate datasets.

### Challenges and Solutions

- **Memory Issues:** Memory limitations were addressed by processing taxi data in chunks.
- **Data Type Mismatches:** The date format was standardized to resolve merging issues.

## Result and Limitations

### Results

- Two separate datasets for 2019 were generated: one for weather data and one for taxi data.
- The Taxi Data collection includes daily aggregated metrics such as total passenger counts, total trips, and average journey duration (in minutes).
- The Weather Data dataset provides daily weather measurements, including average humidity, wind speed, visibility, and maximum and minimum temperatures.

### Data Structure and Quality

- Both datasets are structured compactly, comprising 365 rows (one for each day of the year) with relevant columns optimized for analysis.
- Missing weather data was handled using forward-filling, and the date column was standardized for consistency.

### Limitations

- **Sampling Bias:** Taxi data only includes yellow taxis, excluding rideshare services and green taxis, potentially introducing bias.
- **Data Accuracy:** In addition, both datasets depend on the source providers' correctness and completeness, which may not always be accurate or complete.