Data Analysis Programming Project

NYC TLC Yellow Taxi — 2021 Dataset

1 Project information

- **Project Title:** NYC TLC Yellow Taxi Trip Data Analysis (2021)
- **Group Name:** 4GB RAM
- **Commit hash:** eaa586d498820911fd6da19aa5669e7a70ae59fc
- **Git tag:** v1.0-final
- **Members:** Phạm Quang Minh (24022410), Phạm Nguyễn Xuân Tùng (24022488) , Phạm Vũ Nam(24022416).
- **Data Year:** 2021
- **Submission Date:** 19th December 2025

2 Introduction and Data Overview

2.1 Dataset Context

The project utilizes the **Yellow Taxi Trip Records dataset** provided by the New York City Taxi and Limousine Commission (TLC). This dataset is a standard benchmark for large-scale data analysis, containing detailed records of taxi trips in New York City.

The specific period under analysis is the full year of 2021. This timeframe is particularly significant as it represents a period of post-pandemic recovery, where urban mobility patterns and taxi demand were gradually stabilizing following the disruptions of COVID-19.

2.2 Data Structure and Schema

The raw data is distributed in Apache Parquet format, partitioned by month (12 files from January to December). Each record represents a single taxi trip and contains 19 features.

2.2.1 Main Trip Data

The primary dataset includes the following key attributes:

- **Temporal:** `tpep_pickup_datetime`, `tpep_dropoff_datetime`.
- **Spatial:** `PULocationID` (Pickup), `DOLocationID` (Dropoff).
- **Trip Details:** `passenger_count`, `trip_distance`, `RatecodeID`, `store_and_fwd_flag`.
- **Financial:** `fare_amount`, `extra`, `mta_tax`, `tip_amount`, `tolls_amount`, `improvement_surcharge`, `congestion_surcharg` `airport_fee`, and `total_amount`.
- **Metadata:** `VendorID`, `payment_type`.

2.2.2 Auxiliary Data: Taxi Zone Lookup

Since the trip data uses numerical IDs for locations, an auxiliary CSV file (`taxi_zone_lookup.csv`) is used to map these IDs to meaningful geographical names.

- **LocationID:** Unique identifier (matches PULocationID/DOLocationID).
- **Borough:** The broad administrative district (e.g., Manhattan, Queens).
- **Zone:** The specific neighborhood name (e.g., JFK Airport, Times Sq).

2.3 Initial Data Observations

Preliminary exploration via the `1_view_data.ipynb` notebook revealed several data characteristics and quality issues requiring intervention:

1. **Data Types:** The raw schema consists of a mix of timestamps, integers (IDs, counts), and floating-point numbers (fares, distances).
2. **Value Ranges:**
   - *Fares & Distances:* Values are generally positive, but negative entries (likely refunds or errors) and zero values were observed.
   - *Time:* Most trips fall within 2021, but some outliers exist with years outside the target range.
3. **Missing Values:** Certain columns, such as `passenger_count` and `RatecodeID`, contain null values due to data capture limitations in specific taxi meters.
4. **Categorical Codes:** Fields like `RatecodeID` (1-6) and `payment_type` (1-6) are numerical codes that require mapping to their descriptive string labels for analysis.

These observations form the basis for the cleaning and normalization strategies detailed in the subsequent section.

3 Data Cleaning and Standardization

To ensure analytical accuracy, the raw data from 2021 underwent a rigorous cleaning and standardization process based on defined business rules.

3.1 Data Cleaning

The data cleaning process utilizes a set of rule-based quality checks. Each rule generates a specific quality-assurance flag. Depending on the severity of the violation, records are either excluded from the analysis or flagged for further inspection.

Rule 1: Duplicated rows This rule identifies records that are exact duplicates of previously occurring rows. Duplicate records are removed to prevent double-counting and analytical bias.

RULES RELATED TO DATETIME AND TRIP LOGIC The following rules ensure the temporal validity and logical consistency of the trips.

- **Missing pickup or dropoff datetime (Exclude):** Records missing pickup or dropoff timestamps are excluded, as trip duration and chronology cannot be accurately determined.
- **Dropoff before pickup (Exclude):** Records where the dropoff time precedes the pickup time violate basic causality and are excluded.
- **Invalid pickup month or year (Exclude):** Records with pickup times outside the target analysis month or year are excluded to ensure temporal consistency.

RULES RELATED TO TRIP FEATURES These rules verify the physical plausibility of trip attributes.

- **Non-positive trip duration (Exclude):** Trips with zero or negative duration are excluded as they do not represent valid trips.
- **Non-positive trip distance (Exclude):** Trips with zero or negative distance are excluded due to physical implausibility.
- **Non-positive average speed (Exclude):** Trips with zero or negative average speed are excluded, indicating data errors or invalid records.
- **Zero fare with positive distance (Flag):** Trips with a fare of zero but a travel distance greater than zero are flagged as anomalies.
- **Short duration with long distance (Flag):** Trips with a duration under 1 minute but a distance exceeding 1 mile are flagged due to unrealistic travel speeds.

- **Excessive average speed (Flag):** Trips with average speeds exceeding 66 mph are flagged as inconsistent with typical urban traffic conditions.
- **Excessive trip duration (Flag):** Trips lasting longer than 24 hours are flagged as potential data recording errors.

RULES RELATED TO PAYMENT AND AMOUNTS These rules check the validity and arithmetic consistency of payment and cost information.

- **Non-positive fare amount (Flag):** Trips with a zero or negative base fare are flagged as invalid or erroneously recorded.
- **Negative tip amount (Flag):** Trips with negative tip amounts are flagged as they do not exist in standard transactions.
- **Negative extra charges (Flag):** Trips with negative extra charges are flagged due to arithmetic inconsistency.
- **Negative tolls amount (Flag):** Trips with negative toll fees are flagged as invalid.
- **Non-positive total amount (Flag):** Trips with a zero or negative total payment are flagged, as the total cost must be positive.
- **Fare-total arithmetic mismatch (Flag):** Trips where the recorded total amount deviates significantly from the calculated sum of components are flagged, indicating potential calculation or entry errors.
- **Invalid payment type (Flag):** Trips with a payment type code outside the valid range are flagged as invalid.
- **Invalid rate code (Flag):** Trips with a rate code not belonging to the allowable set are flagged due to potential coding errors.
- **Unusual passenger count (Flag):** Trips with a passenger count of zero or excessively high numbers are flagged as anomalies.
- **Missing pickup or dropoff zone information (Flag):** Trips lacking zone information for either pickup or dropoff are flagged due to incomplete spatial data.

3.2 Data Normalization

The normalization step transforms raw data into a consistent, interpretable structure ready for Quality Assurance (QA) and subsequent analysis. This process focuses on formatting, semantic enrichment, and feature engineering, without removing records.

Datetime normalization The columns `tpep_pickup_datetime` and `tpep_dropoff_datetime` are converted to `datetime` objects and localized to the `America/New_York` timezone. Invalid, ambiguous, or nonexistent values are coerced to `NaT` to ensure consistency when calculating trip duration.

Categorical normalization Numeric categorical variables are mapped to meaningful descriptive labels: `RatecodeID` is mapped to `ratecodeID_name`, and `payment_type` is mapped to `payment_type_name`. Original codes are retained for validation purposes.

Spatial normalization Pickup and dropoff location information is enriched by merging `PULocationID` and `DOLocationID` with the taxi zone lookup table. This process generates semantic columns including `PU_Borough`, `PU_Zone`, `DO_Borough`, and `DO_Zone`, converting location codes into meaningful spatial data.

Trip feature construction New trip features are derived from time and distance data. Specifically, trip duration is calculated in seconds (`trip_duration_seconds`) and rounded to minutes (`trip_duration_minutes`). Average speed (`avg_speed_mph`) is derived from distance and duration, where infinite values caused by division by zero are handled as `NaN`.

Temporal features Based on the pickup time, additional temporal features are generated, including `pickup_day_of_week` and an `is_weekend` flag, to support cyclical behavioral analysis.

Fare consistency feature A calculated total amount variable (`computed_total_amount`) is created by summing known cost components, with missing values replaced by 0. This variable is not used to alter data but serves as a reference for checking consistency against the recorded total amount during the QA step.

Column reduction Several columns are removed from the dataset as they are irrelevant to the analysis objectives, exhibit low variance, or contain excessive missing values. These include `airport_fee`, `store_and_fwd_flag`, `VendorID`, `mta_tax`, and `improvement_surcharge`.

3.3 Quality Assurance Summary

A comprehensive audit of the 2021 dataset was conducted using the defined QA rules. The summary of violations reveals distinct patterns in data quality, ranging from systemic arithmetic inconsistencies to rare anomalies.

3.3.1 Violation Statistics

Table 1 highlights the rules with the highest violation rates averaged across 12 months. Notably, approximately **39%** of all records were flagged by at least one QA rule, necessitating a cautious approach to data filtering (using soft flags instead of hard exclusions for non-critical errors).

ID	Rule Name	Action	Avg. Rate (%)	Hypothesized Cause
17	Fare-Total Arithmetic Mismatch	Flag	$\sim$32.5%	Taximeter rounding errors or undocumented surcharges.
19	Invalid RatecodeID	Flag	$\sim$5.1%	Driver manual entry error or system default settings.
20	Unusual Passenger Count (0 or > 5)	Flag	$\sim$3.8%	Group rides or failure to reset passenger count.
21	Invalid Zone ID	Flag	$\sim$1.7%	GPS drift or map version mismatch.
7	Invalid Speed (≤ 0)	Exclude	$\sim$1.4%	GPS signal loss in tunnels or stationary idling.

Table 1: Top Data Quality Violations (Averaged for 2021)

3.3.2 Data Quality Insights

Based on the violation distribution, we derived the following observations regarding the reliability of the NYC TLC dataset:

- **Systemic Fare Inconsistencies:** The `fare_total_mismatch` rule triggered the highest volume of flags (peaking at 35.1% in November). This pervasive issue suggests a systemic discrepancy between the recorded `total_amount` and the sum of individual components (fare, tax, tolls). This is likely attributed to legacy taximeters calculating totals differently than the server-side aggregation logic, or small rounding differences. As these are valid trips, they are flagged rather than excluded to preserve data volume.
- **Sensor and Human Input Errors:** Physical impossibility rules, such as `invalid_distance` and `invalid_speed`, consistently flagged about 1.2%–1.9% of records. These are characteristic of GPS jitter or signal loss (common in NYC's "urban canyons"). Meanwhile, `invalid_ratecode` and `unusual_passenger_count` errors suggest manual input fatigue by drivers.
- **High Reliability Fields:** Critical transactional fields such as `tpep_pickup_datetime`, `payment_type`, and `tip_amount` exhibited negligible error rates ($< 0.01\%$). This indicates that the automated digital logging systems for time and payments are highly robust and reliable for downstream analysis.
- **Seasonal Stability:** The total violation rate remained relatively stable around 38%–39% throughout the year, implying that data quality issues are structural rather than seasonal or event-driven.

A significant finding from the QA process is the exceptional reliability of transactional and temporal data fields, where violation rates consistently remained below **1%**.

- **Transactional Robustness:** Financial inputs such as `invalid_payment_type` (0.0%), `invalid_tip_amount` ($\sim$0.01%), and `invalid_tolls_amount` ($\sim$0.01%), showed negligible error rates. This strongly suggests that financial data is captured via automated digital payment gateways rather than manual entry, ensuring high precision for revenue analysis.

- **Temporal Consistency:** The `missing_datetime` and `invalid_month` rules yielded near-zero violations. This indicates that the On-Board Diagnostic (OBD) systems in taxis function reliably in timestamping trips, with very few gaps or synchronization errors.
- **Data Uniqueness:** The `is_duplicate` rule flagged almost 0 records throughout the year. This confirms the efficacy of the data ingestion pipeline in maintaining record uniqueness, eliminating the need for aggressive de-duplication steps during preprocessing.

Conclusion: The low variance in these specific error rates validates the dataset's suitability for time-series forecasting and financial modeling, as the core variables (Time and Money) are highly accurate.

4 Aggregation and KPI Definition

Following the cleaning and normalization stages, data is aggregated into Daily, Weekly, and Monthly timeframes. Key Performance Indicators (KPIs) are defined to analyze trip behavior, spatial dynamics, speed, and revenue. Specific metrics incorporate Quality Assurance (QA) filters to ensure calculation accuracy by excluding invalid records.

4.1 Key Performance Indicators (KPIs)

KPI Name	**Significance & Objective**	**Formula**	**Unit**
Total Trips	Total number of trips per period, reflecting overall demand.	$N = \sum_{i=1}^{n} 1$	Trips
Duration p50	Typical trip duration (median), robust against outliers.	$Q_{0.5}(D_i)$	Minutes
Duration p95	Unusually long trip duration, indicative of potential congestion.	$Q_{0.95}(D_i)$	Minutes
Speed p50	Typical travel speed (median).	$Q_{0.5}(S_i)$	Mph
Distance p50	Typical trip distance (median).	$Q_{0.5}(L_i)$	Miles
Distance p95	Long-distance trips (95th percentile), often reflecting suburban routes.	$Q_{0.95}(L_i)$	Miles
Avg Distance	Mean trip distance.	$\bar{L} = \frac{1}{n}\sum_{i=1}^{n} L_i$	Miles
Revenue per Trip	Average revenue per valid trip.	$R_{trip} = \frac{\sum_{i \in V} A_i}{\lvert V \rvert}$	USD ($)
Revenue per Mile	Revenue efficiency per mile traveled.	$R_{mile} = \frac{\sum_{i \in V} A_i}{\sum_{i \in V} L_i}$	USD/Mile
Trips per Hour (Pickup / Dropoff)	Trip intensity per hour within specific time bins, separated by pickup and dropoff directions.	$\frac{\sum I_{i,b}^{PU}}{H_b}$ / $\frac{\sum I_{i,b}^{DO}}{H_b}$	Trips/Hour

Table 2: KPI Definitions, Formulas, and Units

4.2 Index-based Normalization Metrics

In addition to absolute KPIs, this study employs index-based metrics to normalize time-series data. This approach facilitates the comparison of fluctuation patterns across days and weeks, allowing for the analysis of relative trends while mitigating the effects of absolute scale and short-term seasonality.

Daily Index (Index_100_by_Day). This metric normalizes daily KPI values relative to the first day of the time series, setting the baseline value to 100. This method highlights relative growth or decline trends over time.

$$\text{Index}_t^{Day} = \frac{X_t}{X_{t_0}} \times 100$$

Where:

- X_t is the KPI value at day t,
- X_{t_0} is the KPI value of the first day in the observed series.

5 Data Visualization

***** *This section presents a set of visualizations to explore temporal, behavioral, and traffic-related characteristics of taxi activity in January 2021. The figures focus on daily demand dynamics, intra-week temporal patterns, traffic conditions, and passenger payment behavior.*

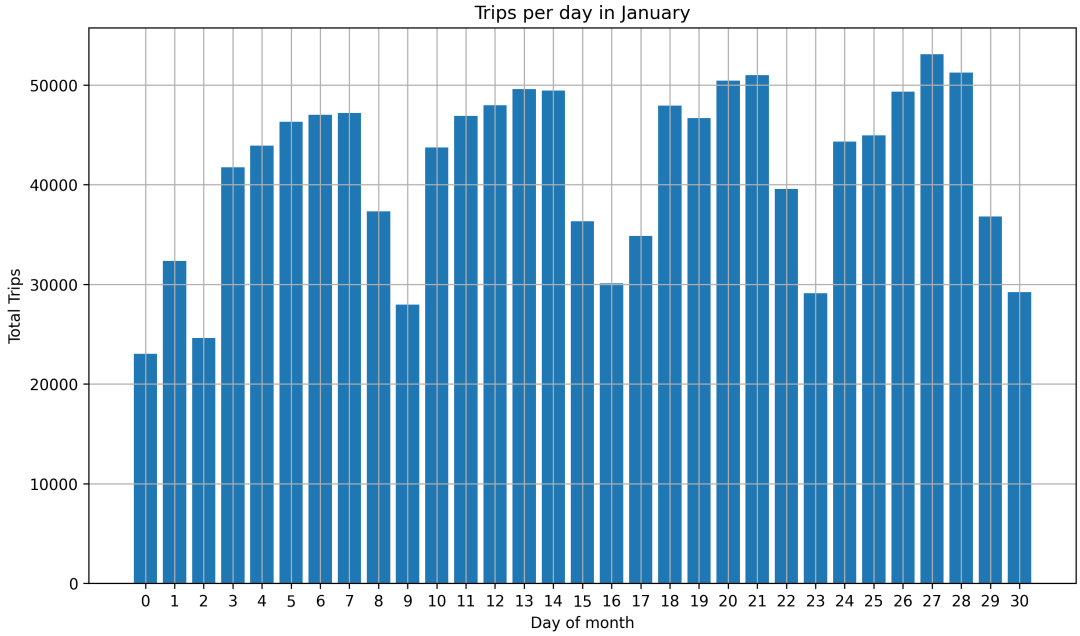


Figure 1: Daily taxi trip volume in January 2021, illustrating day-to-day variability in demand. Noticeable fluctuations are observed throughout the month, with lower activity at the beginning and toward the end..

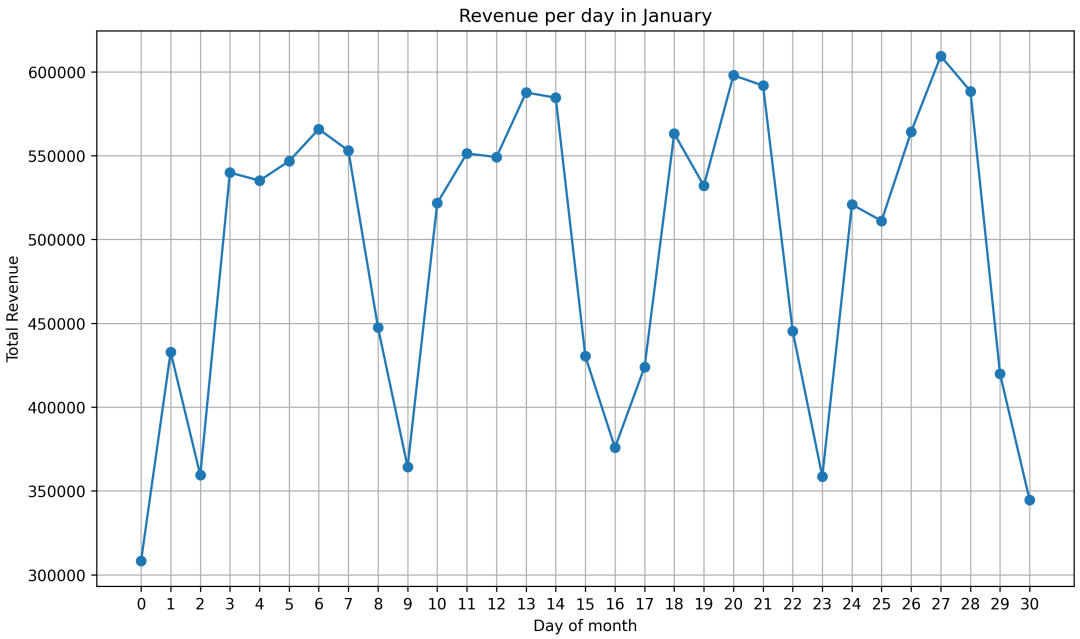


Figure 2: Daily taxi revenue in January 2021, showing short-term fluctuations in total earnings. Revenue patterns closely follow variations in trip volume, with several sharp increases and decreases. Despite occasional peaks, overall revenue remains moderate during this period.

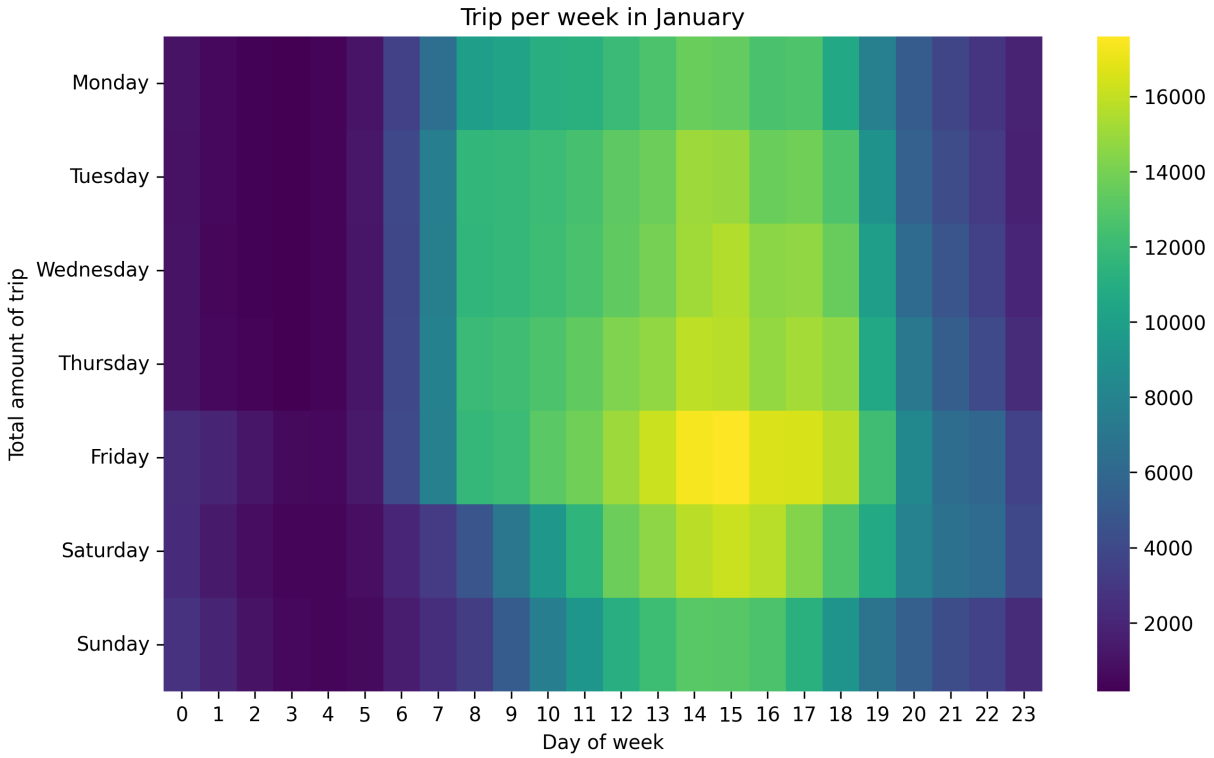


Figure 3: Heatmap of taxi trip volume by day of week and hour of day in January 2021. Higher demand is mostly concentrated during midday and early afternoon hours, particularly on weekdays.

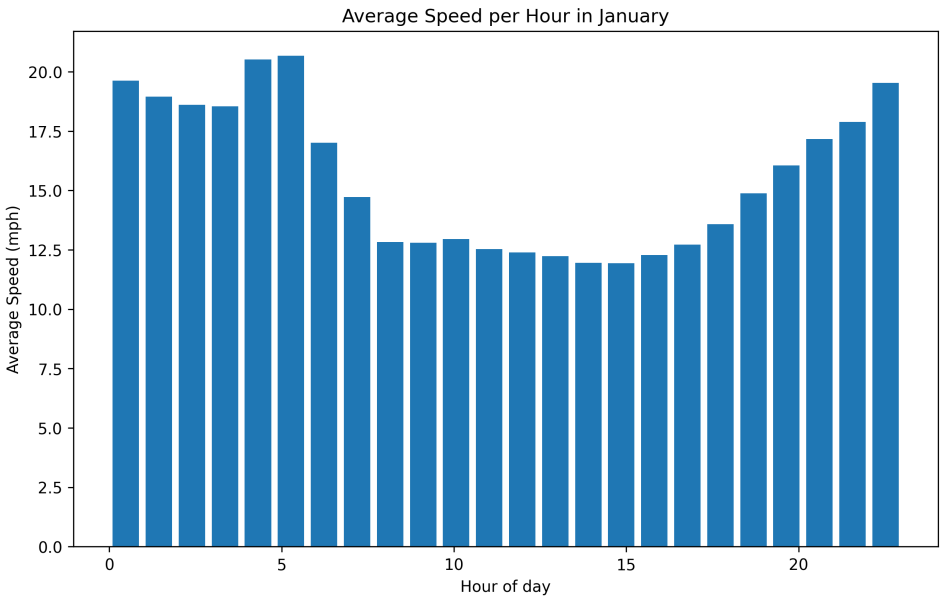


Figure 4: Average taxi speed by hour of day in January 2021. Higher speeds are observed during late-night and early-morning hours, as the traffic are usually less frequent.

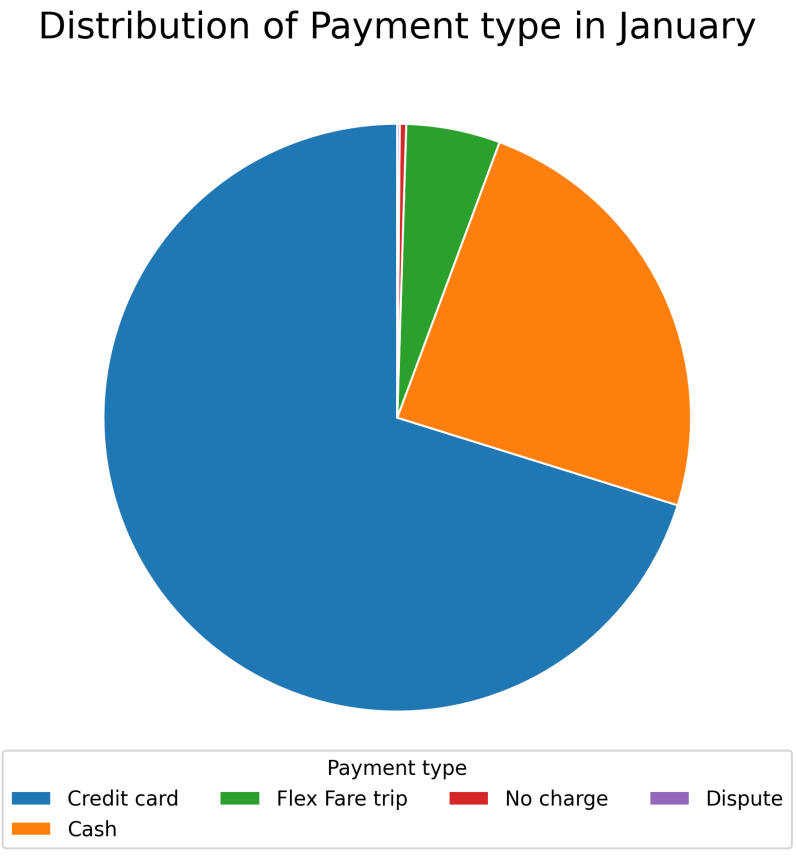


Figure 5: Distribution of taxi payment types in January 2021. Most passengers prefer to pay by credit card, while cash is used less frequently. Some trips are even charged under negotiation or none at all.

Logarithmic Transformation Rationale ($\log(1+x)$) Taxi trip data typically exhibits a heavy right-skewed distribution, where the majority of trips are short, but a long tail of rare, extreme values exists. Direct visualization on a linear scale would compress the dense region of short trips, making patterns indistinguishable. To address this, we apply the $\log(1+x)$ transformation.

This method offers two key advantages:

1. **Handling Zero Values:** The addition of 1 ensures stability for zero-valued entries (since log(0) is undefined), allowing us to handle valid trips with very short distances.

2. **Focus on Relative Magnitude:** The log scale emphasizes proportional differences rather than additive ones. For instance, in operational terms, a trip increasing from **5 to 10 minutes** (a 100% increase) is a significant change in behavior compared to an increase from **55 to 60 minutes** (a ∼9% increase), even though the absolute difference is 5 minutes in both cases. The log transformation normalizes these scales, allowing us to visualize short urban hops and long suburban commutes within the same distribution.

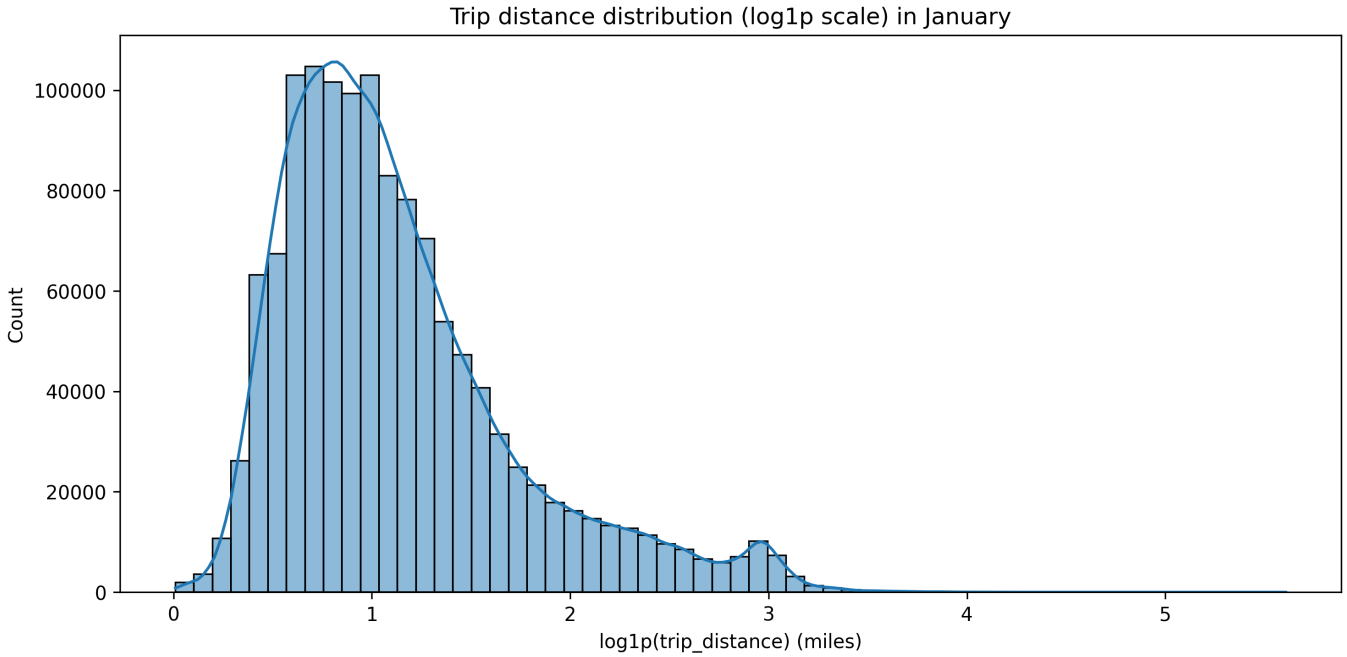


Figure 6: Distribution of trip distances in January 2021

The histogram illustrates the distribution of taxi trip distances after applying a log transformation to reduce skewness. Most trips are relatively short, with a long right tail indicating a small number of long-distance journeys.

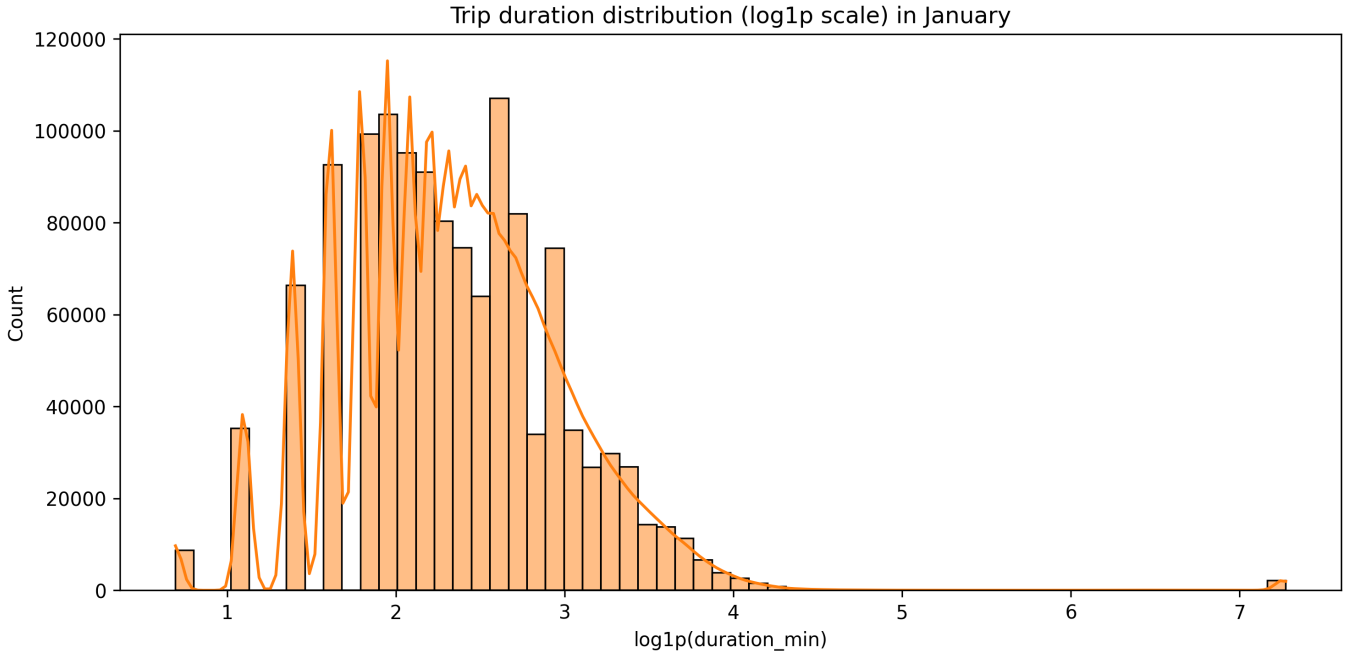


Figure 7: Histogram of trip durations for January
The peak density centers around a log-value of approximately 2.5 to 2.8. Transforming this back ($\exp(x) - 1$), the typical taxi trip lasts between **11 and 15 minutes**. This confirms that the primary use case for Yellow Taxis is short-to-medium distance intra-city travel. The symmetry of the curve suggests that the data cleaning process (Section 3) successfully removed non-physical outliers (e.g., negative time or multi-day trips), leaving a natural and coherent set of trip records for analysis.

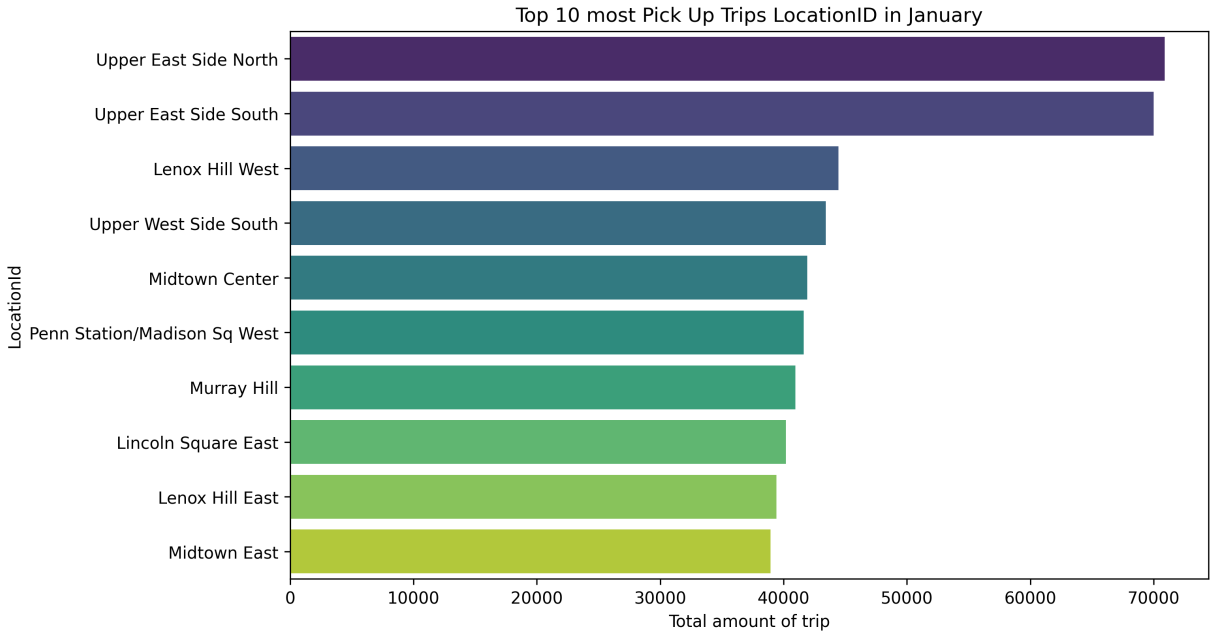


Figure 8: Top 10 pickup LocationIDs in January 2021. The horizontal bar chart shows the most frequent pickup LocationIDs (or zone names when available) ranked by trip count; counts reflect valid pickups after QA filtering and highlight concentrated demand in central zones versus a long tail of less frequent pickup locations.

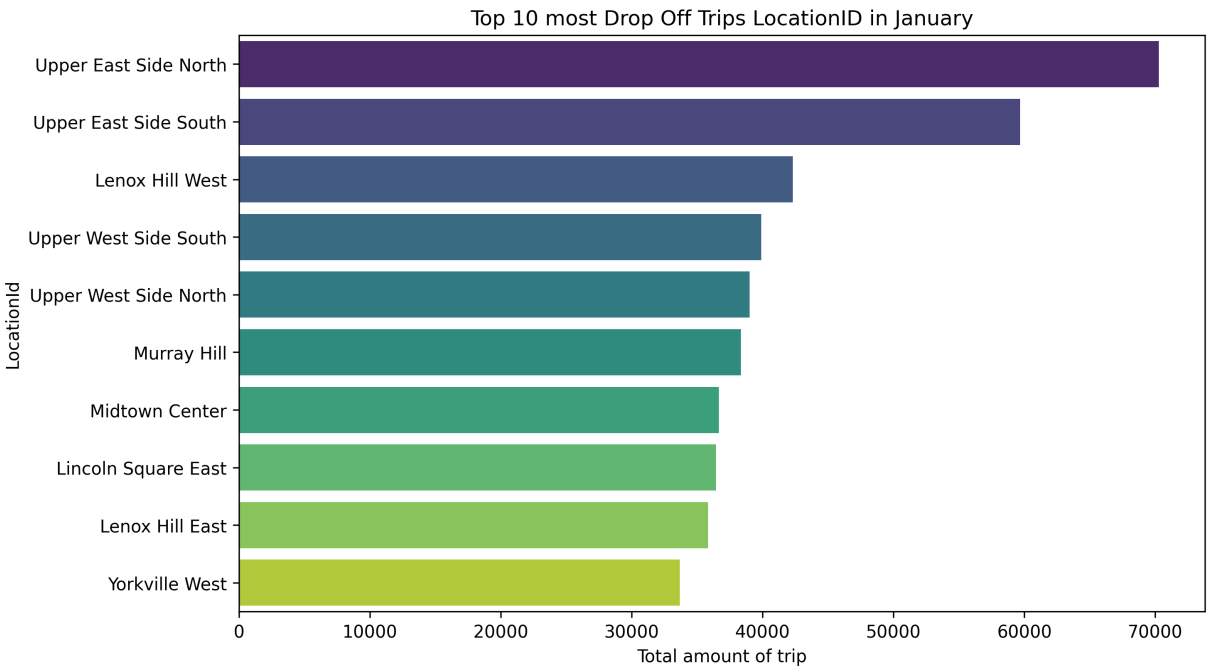


Figure 9: Top 10 dropoff LocationIDs in January 2021. The plot ranks the busiest dropoff LocationIDs by trip count and percentage of valid dropoffs, revealing main destination hotspots and distributional differences compared with pickup locations.

6 Advanced Analysis Models

6.1 Trip Demand Forecasting Algorithm

The primary objective of this algorithm is to construct a predictive model for future taxi demand based on historical data, thereby supporting efficient fleet management and allocation. By benchmarking multiple methodologies, we identify the model best suited for the specific characteristics of taxi trip data while evaluating the performance gains of complex models against a baseline.

To achieve this, we implement and compare three distinct approaches: a Baseline model, an ARIMA model, and a Linear Regression model. Trip data is aggregated into a time series $\{y_t\}$ at the desired frequency (hourly or daily), then split into a training set $\{y_1, ..., y_{n-k}\}$ and a test set $\{y_{n-k+1}, ..., y_n\}$, where k denotes the number of testing periods.

6.1.1 Forecasting Models

Baseline Model (Seasonal Naive)

The baseline model exploits the inherent seasonality of the data by projecting historical values from the same period in the previous cycle:

$$\hat{y}_t = y_{t-s} \tag{1}$$

where s is the seasonal period length ($s = 168$ hours for weekly cycles with hourly data, or $s = 7$ days for daily data). If y_{t-s} is unavailable in the training set, the global mean of the training data is used as a fallback. This model serves as a benchmark for evaluating more complex algorithms.

ARIMA Model

The ARIMA(p, d, q) model is a standard approach for time-series forecasting, integrating three components: Auto-Regressive (AR) of order p, Integrated (I) of order d, and Moving Average (MA) of order q. The model is expressed as:

$$\phi(B)(1 - B)^d y_t = \theta(B)\epsilon_t \tag{2}$$

where B is the backshift operator ($By_t = y_{t-1}$), $\phi(B)$ and $\theta(B)$ are characteristic polynomials, and ϵ_t is white noise. Model parameters are estimated using Maximum Likelihood Estimation (MLE) on the training set.

Linear Regression Model

A simple linear regression model assumes a deterministic linear trend over time:

$$y_t = \beta_0 + \beta_1 \cdot x_t + \epsilon_t \tag{3}$$

where x_t represents the time index (days elapsed since the start), β_0 is the intercept, β_1 is the trend coefficient, and ϵ_t is the error term. Parameters are estimated using the Ordinary Least Squares (OLS) method:

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0,\beta_1} \sum_{t=1}^{n-k} (y_t - \beta_0 - \beta_1 x_t)^2 \tag{4}$$

Despite its simplicity, this model effectively captures clear upward or downward trends in the dataset.

6.1.2 Performance Evaluation

Model performance is evaluated on the test set using three standard metrics. **MAE (Mean Absolute Error)** measures the average absolute deviation:

$$MAE = \frac{1}{k} \sum_{i=1}^{k} |y_i - \hat{y}_i| \tag{5}$$

MAPE (Mean Absolute Percentage Error) provides a relative measure of error, suitable for comparing datasets of varying scales:

$$MAPE = \frac{1}{k} \sum_{i=1}^{k} \frac{|y_i - \hat{y}_i|}{y_i} \times 100\% \tag{6}$$

RMSE (Root Mean Squared Error) penalizes larger errors more heavily:

$$RMSE = \sqrt{\frac{1}{k} \sum_{i=1}^{k} (y_i - \hat{y}_i)^2} \tag{7}$$

Lower values across these metrics indicate superior model performance.

6.2 Spatio-Temporal Zone Clustering Algorithm

The goal of this algorithm is to group zone-time pairs into clusters with similar mobility demand and traffic characteristics. This segmentation identifies high-volume congested zones, smooth-flow low-demand areas, or regions with unstable traffic, facilitating targeted fleet operations.

6.2.1 Feature Engineering and KPI Calculation

Trip data is grouped by pickup zone and time of day. Time is discretized into 6 specific bins reflecting traffic patterns: Early Morning (0-4h), Morning (4-7h), Morning Rush (7-10h), Midday (10-16h), Evening Rush (16-19h), and Late Night (19-24h). For each (zone, time_bin) group, the following features are computed:

- **duration_p50**: Median trip duration, representing typical traffic conditions.
- **duration_p95**: 95th percentile trip duration, capturing congestion or anomalies.
- **speed_p50**: Median trip speed.
- **avg_trip_distance**: Mean trip distance.
- **trips**: Total trip count.
- **trips_index_100**: Normalized trip volume index, calculated as:

$$\text{trips\_index\_100} = \frac{\text{trips}}{\overline{\text{trips}}} \times 100 \tag{8}$$

where $\overline{\text{trips}}$ is the global average trip count across all groups.

6.2.2 K-Means Clustering

We apply the K-Means algorithm to cluster these groups based on three key features: `duration_p50`, `duration_p95`, and `trips_index_100`. Prior to clustering, features are standardized using a StandardScaler:

$$x_{scaled} = \frac{x - \mu}{\sigma} \tag{9}$$

The K-Means algorithm identifies k centroids $\{\mathbf{c}_1, ..., \mathbf{c}_k\}$ by minimizing the within-cluster sum of squares:

$$J = \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} ||\mathbf{x} - \mathbf{c}_i||^2 \tag{10}$$

Optimization proceeds iteratively through Assignment and Update steps until convergence.

6.2.3 Cluster Interpretation and Labeling

Post-clustering, centroids are analyzed to assign semantic labels based on defined business rules. This automated labeling logic is implemented in the `cluster_zone_time` function:

- **High Demand – Congested:** Clusters with `trips_index_100` > 500 and `duration_p95` > 20 min, indicating high volume with significant delays.
- **Low Demand – Smooth Flow:** Clusters with `trips_index_100` < 50 and `duration_p50` < 30 min, indicating low activity and free-flow traffic.
- **Unstable Traffic:** Clusters with `duration_p95` > 50 min, suggesting highly variable travel times.
- **Efficient High Volume:** Remaining clusters representing high demand but reasonable travel durations (efficient operations).

This labeling strategy transforms raw clusters into actionable insights, characterizing the operational state of each zone at specific times of the day.

7 Business Insights and Strategic Recommendations

7.1 Key Findings

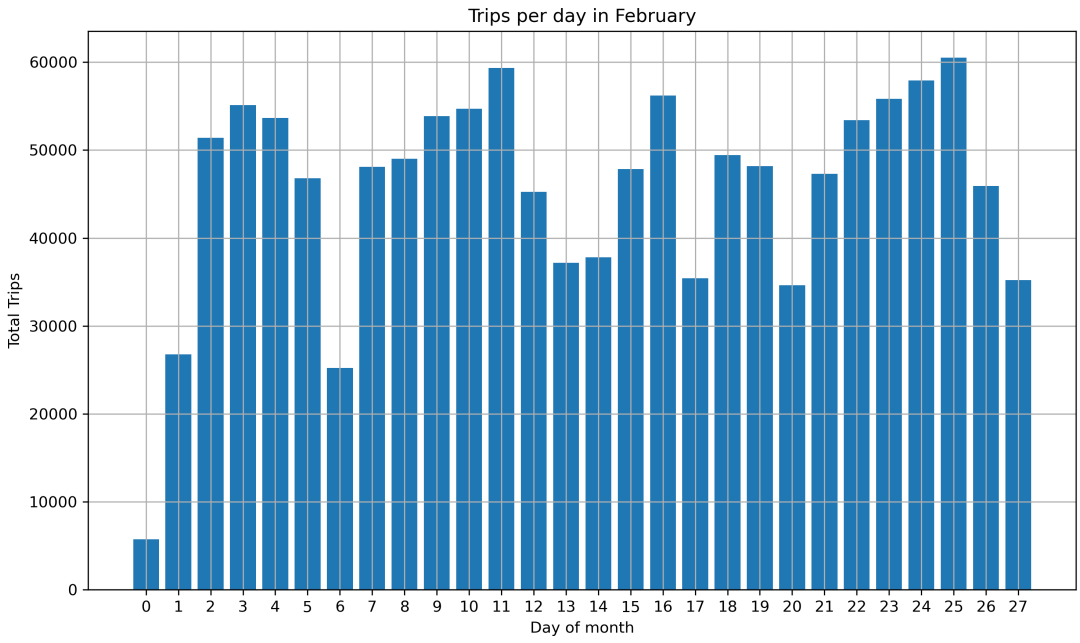


Figure 10: February daily trips, with somewhat higher trips than January

- **Early 2021 (January–February)** shows persistently low levels of trips and revenue (see figure 1 and 10), reflecting the peak impact of COVID-19 restrictions, remote work policies, and public risk aversion. Trips in

the earlier months tends to distribute more on the working hours, usually from 7 AM to 8 PM especially on the working days (see figure 3).

- A structural inflection point emerges around **March-April 2021**, where both trip volume and revenue begin to accelerate. This period marks the transition from stagnation to recovery. "Late night trips" models are also starting to appear in May since people are ordering trips up until 2 AM. From here there were no signs of halt in both trips and revenue are gradually increasing (and almost double the total trips of January!).

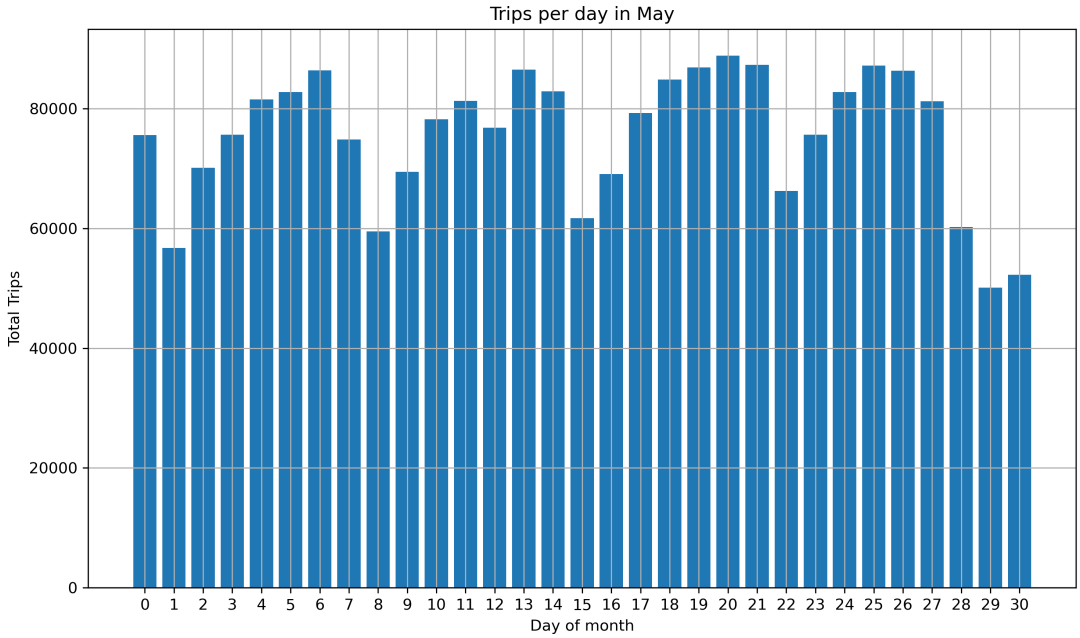
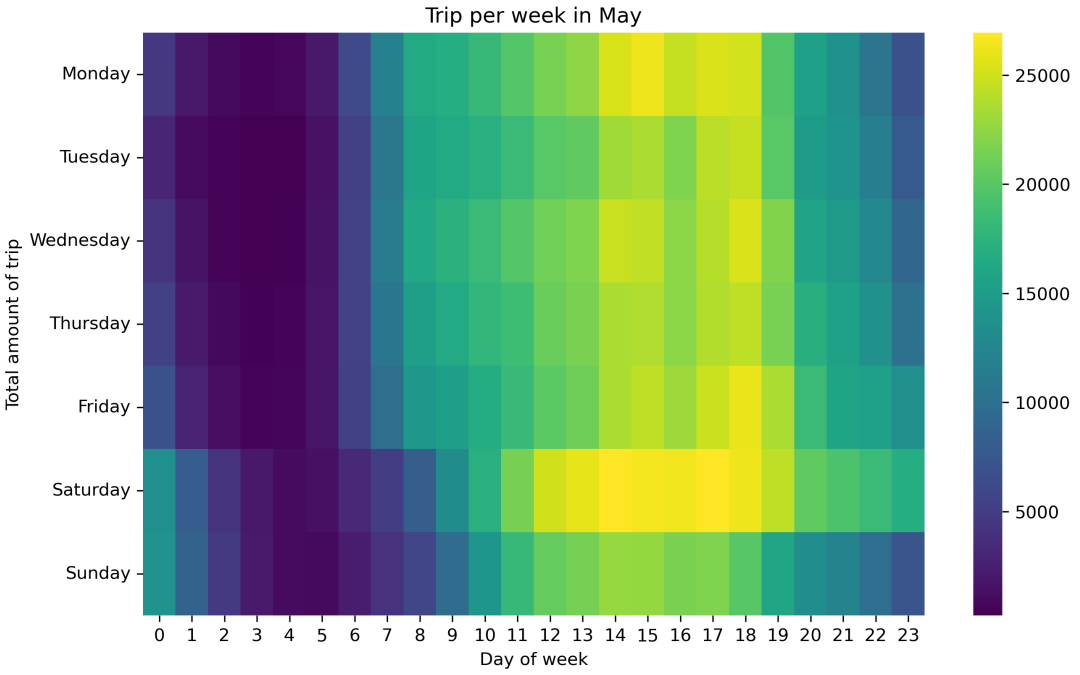


Figure 11: May daily trips



- **From June onward**, demand still grows steadily, with minor fluctuations during mid-summer, and reaches its highest levels in July-December, coinciding with the reopening of tourism and large-scale social activities, like the Independence Day, Halloween, Thanksgiving or Christmas. Revenues still runs along with the schedule of the week but with the exception of December with a sudden drop between December 22nd–24th due to the public staying home for Christmas observances.

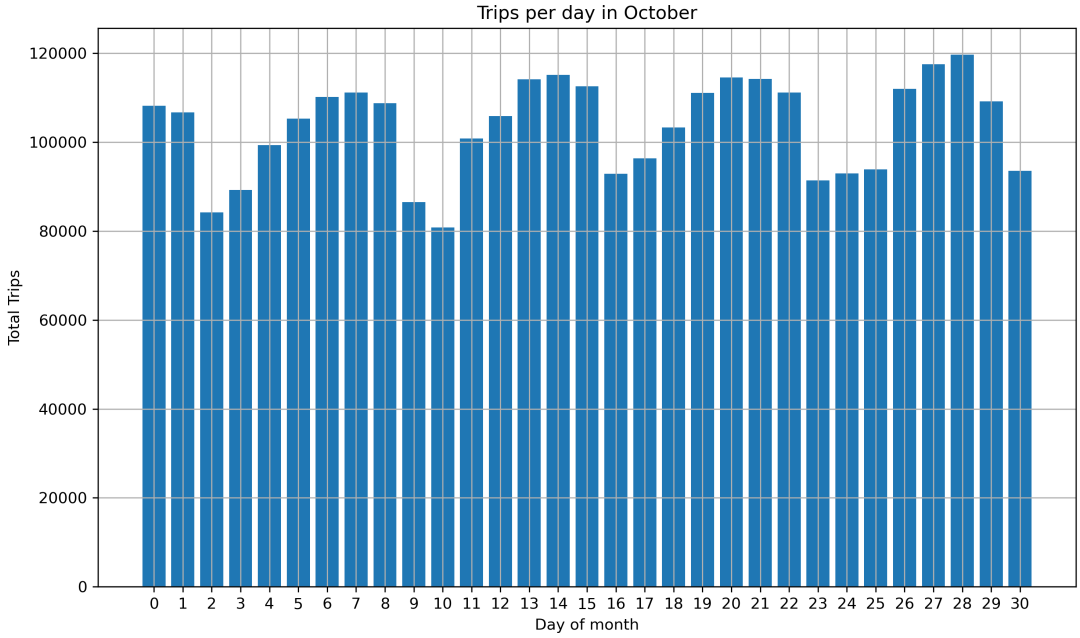


Figure 12: October daily trips

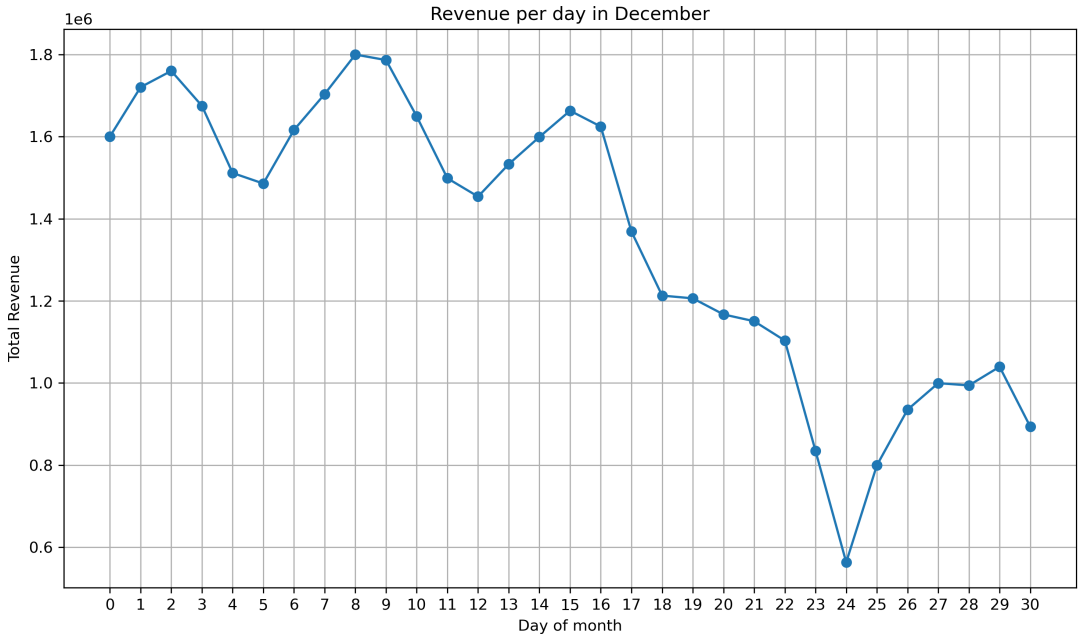


Figure 13: December daily revenue

- Both Upper East Side South and Upper East Side North remains the largest throughout the whole year. With that, there is a noticeable rise in pick up rates of JFK Airport during the year.

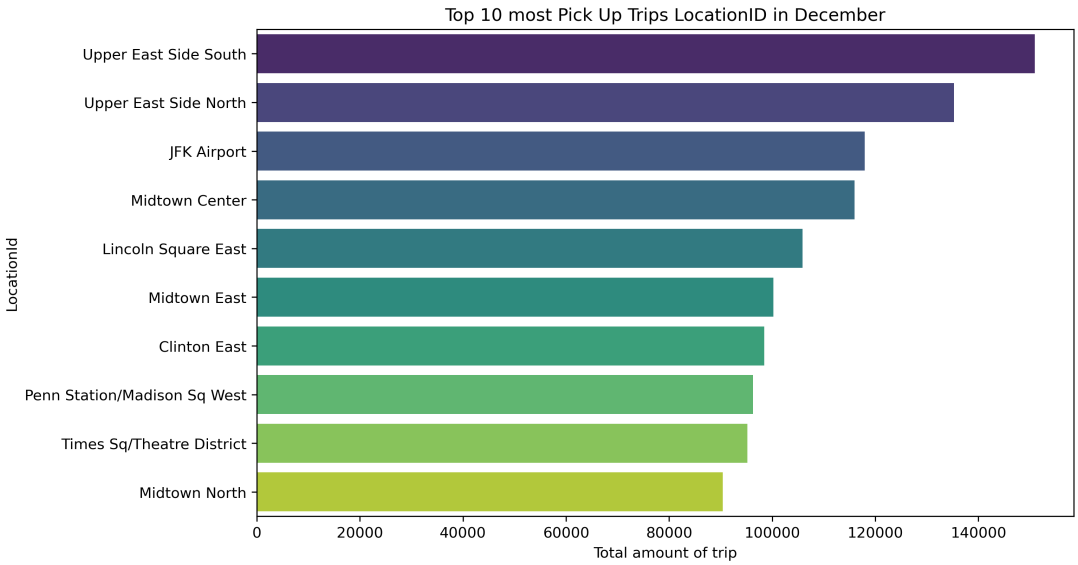


Figure 14: Top 10 PU zones in December

7.2 Context

These trends must be interpreted within the broader public health and economic context of New York City in 2021.

- During the first months of 2021, NYC is undergoing serious COVID-19 pandemic, with high infection rates, strict indoor activity limits, and suppressed mobility. This is directly causing poor performance in taxi drivers, although vaccine is being administered.
- From February to March, public schools began phased reopening and cluster-zone restrictions were lifted, signaling the first steps toward normalization. This explains the gradualling increase in all aspects from this month onward, including more taxi drivers are back again because of the vaccine. But there were also a snowstorm in February that caused short-term drops in trip volume and occasional revenue spikes due to scarcity and urgency.
- The restoration of 24-hour subway service and the city's full reopening in mid-2021 enabled the recovery of nightlife, leisure, and airport-related travel, key revenue sources for yellow taxis.
- By June of 2021, the city's overall testing positivity rate had reached its lowest since the pandemic began. Since the first vaccines arrived in December, over 8,408,000 doses were administered. Governor Cuomo reopened the entirety of New York State on June 15, two weeks ahead of Mayor Bill de Blasio's planned July 1 reopening.
- The emergence of the Delta variant led to renewed health measures but did not materially reverse mobility trends; vaccine mandates helped sustain indoor economic activity and stabilized taxi demand.
- During October–November 2021, enforcement of vaccination mandates for city employees and the holiday season drove a surge in travel and shopping activity, resulting in peak taxi demand in Q4, with particularly strong late-night, weekend, and airport traffic.

7.3 Strategic Recommendations

Based on the observed demand recovery patterns and behavioral shifts throughout 2021, the following strategic recommendations are proposed:

- **Marketing and Demand Stimulation**: Marketing expenditure should be concentrated between **July and December**, when demand momentum is strongest and customer responsiveness is highest. Promotional campaigns during this period can more effectively capture returning tourists, leisure travelers, and late-night riders.
- **Weekday Fleet Allocation**: On weekdays, vehicle supply should be prioritized in central business districts, particularly Manhattan, during peak commuting hours. The re-emergence of the morning and evening *Twin Peaks* suggests renewed demand from office-related travel.
- **Weekend and Nighttime Operations**: During weekends, fleet deployment should shift toward entertainment, dining, and nightlife areas. Incentivizing drivers to operate during late-night hours can help capture high-value leisure demand and benefit from reduced traffic congestion.

- **Weather-Responsive Pricing and Safety Measures**: During extreme weather events such as snowstorms, proactive driver alerts and dynamic pricing (e.g., surge pricing) should be implemented to compensate for increased operational risk while ensuring service availability.

Overall, these recommendations emphasize aligning marketing investment and operational decisions with the temporal and behavioral demand patterns revealed by the data, enabling more efficient resource utilization and improved revenue performance.

8 References

Main datasets: https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.pagences

References

[1] McKinney, W. (2022). *Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter*. 3rd ed., O'Reilly Media. Available at: `https://www.oreilly.com/library/view/python-for-data/9781098104023/`

[2] Timeline of the COVID-19 pandemic in New York City. Wikipedia. Available at: `https://en.wikipedia.org/wiki/Timeline_of_the_COVID-19_pandemic_in_New_York_City`

[3] Record-breaking wintry weather of February 2021. AccuWeather. Available at: `https://www.accuweather.com/en/weather-news/the-record-breaking-wintry-weather-of-february-2021/`