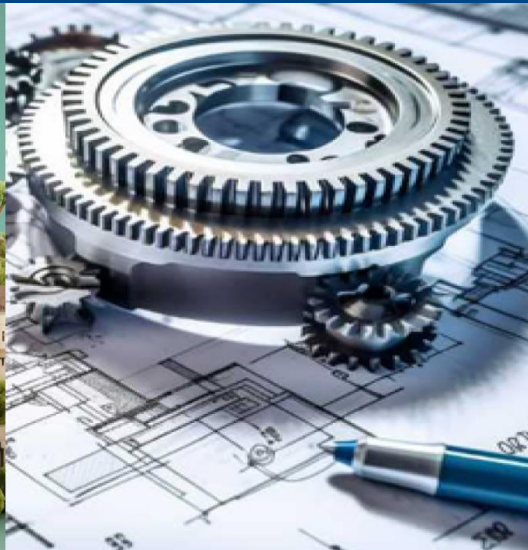
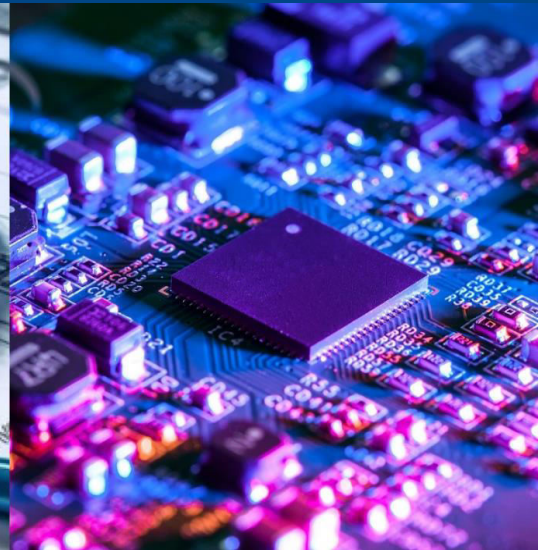
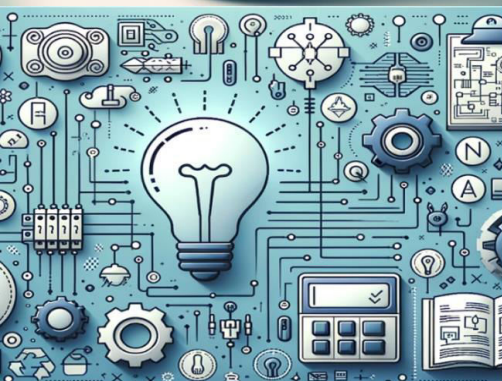




International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.864

Volume 9, Issue 10, October 2026



International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Simulation-Based Analysis of Ride-Hailing Earnings Fairness in Mumbai Using Synthetic Data and NYC TLC Benchmarking

Kaushal Parsekar, Dr. Abuzar Ansari

M.Sc. Student, Department of Data Science, SIES College of Arts, Science and Commerce (Empowered Autonomous),
Sion (West), Mumbai, Maharashtra, India

HOD, Department of Data Science, SIES College of Arts, Science and Commerce (Empowered Autonomous), Sion
(West), Mumbai, Maharashtra, India

ABSTRACT: This study examines January 2025 New York City Taxi and Limousine Commission high-volume for-hire vehicle trip records and separately evaluates earnings-aware dispatch in a synthetic Mumbai-inspired scenario. The NYC analysis summarizes trip-level payments by time and pickup area; it does not estimate driver-level earnings. A five-seed simulation compares a nearest-eligible-driver baseline with a policy that also considers accumulated earnings. The fairness-aware policy lowers simulated driver-level earnings inequality and reduces drivers receiving no trips, while slightly lowering service rate and increasing pickup time. Because Mumbai inputs are synthetic, results are illustrative rather than estimates of actual local driver outcomes.

KEYWORDS: ride-hailing, earnings fairness, Gini coefficient, dispatch simulation, synthetic data.

I. INTRODUCTION

Ride-hailing services allocate passenger requests among drivers whose locations and availability change over time. Assignment choices can influence waiting time, completed trips, geographic coverage, and the distribution of work. Selecting the closest eligible driver may favour fast pickups, whereas including accumulated earnings in the decision can change which drivers receive opportunities. These objectives can conflict, and the outcome depends on demand, driver availability, travel-time assumptions, and the chosen definition of fairness. Existing studies have treated fairness as a relevant part of ride-hailing assignment and optimization [1], [3], [4].

The study is organized into two separate analyses. The first describes trip-level payments in the publicly available January 2025 NYC TLC HVFHV records. The second compares two dispatch rules in a synthetic environment inspired by selected Mumbai locations. The NYC data are used only as an external descriptive benchmark; they are not combined with the synthetic records and are not assumed to represent Mumbai conditions. The source dataset is published by the New York City Taxi and Limousine Commission (TLC) [5].

Research questions

- RQ1: How do recorded positive driver payments per trip vary across time periods and pickup locations in the January 2025 NYC TLC HVFHV data?
- RQ2: Under explicit synthetic Mumbai-inspired assumptions, how does an earnings-aware allocation policy compare with a nearest-available-driver baseline on simulated earnings inequality and service-efficiency metrics?

The paper makes three bounded contributions: it summarizes how NYC trip-level payments differ across time and pickup zones; it sets out a reproducible synthetic Mumbai-inspired scenario for comparing a nearest-driver rule with an earnings-aware heuristic; and it reports the resulting fairness and service metrics across multiple random seeds and selected parameter settings. These experiments are exploratory. They neither measure actual earnings in Mumbai nor demonstrate causal discrimination by a platform.



International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

- RQ3: Do the observed simulation trade-offs persist across random seeds, fairness-weight settings, pickup-time assumptions, and exploratory demand scenarios?

II. LITERATURE SURVEY

Prior work on ride-hailing has examined assignment quality, unequal outcomes, and longer-term effects on drivers. Shi et al. present a fair task-assignment approach for large ride-hailing systems [1]. Bokányi and Hannák investigate service inequality through simulation [2]. Sun et al. consider efficiency and fairness together by coordinating order dispatch with driver repositioning [3]. Tzur and Ben-Gal examine data-driven policies for online ride-hailing with fairness objectives [4]. This literature supports assessing distributional outcomes alongside operational indicators instead of judging a dispatch policy by a single performance measure.

This project takes a deliberately practical approach: it compares an interpretable allocation rule with a nearest-eligible-driver baseline in a repeatable synthetic setting, and separately describes payment patterns in public NYC trip records. It does not claim that the heuristic is unprecedented, that the simulation proves discriminatory treatment, or that the NYC data establish earnings conditions in Mumbai. The cited studies provide background, but a more systematic literature review would be needed to substantiate any strong claim of novelty [1]–[4].

III. METHODOLOGY / APPROACH

A. NYC TLC empirical dataset

The analysis uses the official NYC TLC HVFHV trip-record file for January 2025 [5]. The file contains 20,405,666 records. Fields used include request, pickup and drop-off times, pickup and drop-off zone identifiers, trip distance, trip duration, passenger fare, and driver pay. The notebook reports no missing values in the driver-pay field, alongside 313 records with negative pay and 4,225 with zero pay. Calculations of positive trip payments retain records where driver_pay is greater than zero. The negative and zero values are reported rather than automatically treated as data errors. TLC cautions that records are submitted by for-hire vehicle bases and may not be fully accurate or complete [5].

Summary statistics are calculated by pickup hour, broad time period, weekday/weekend status, and pickup zone. A Gini coefficient is also calculated for positive trip-payment amounts. Its unit of analysis is an individual trip, not a driver. Since this analysis does not have a stable driver identifier together with complete online-time information, the NYC data are not used to calculate actual driver-level hourly earnings or driver-level earnings inequality [5].

B. Synthetic Mumbai-inspired simulation

The synthetic environment represents ten selected areas—CSMT, Dadar, Bandra, BKC, Andheri, Kurla, Powai, Ghatkopar, Thane, and Kalyan. Their coordinates are simplified grid positions, not geographic coordinates or road-network distances. Demand weights, hourly request volumes, driver locations and shifts, trip distances and durations, and fare/pay values are generated from stated assumptions. The baseline scenario includes 500 synthetic drivers and 6,320 requests. Simulated pay follows a distance-and-duration rule with random variation and a minimum amount. None of these generated values should be interpreted as observed market data from Mumbai.

C. Synthetic Mumbai data construction and assumptions

No field survey or platform dataset was collected for the Mumbai component. Instead, the simulation generates a controlled test environment so that the two assignment rules can be compared under identical conditions. The notebook specifies ten areas, relative pickup weights, illustrative coordinates, and a 24-hour request profile, then creates driver schedules and ride requests using a seeded random-number generator. For each seed, the baseline and fairness-aware strategies are tested against the same generated scenario, making the comparison more controlled and reproducible.

Zones and assumed relative pickup weights

CSMT: 0.08; Dadar: 0.10; Bandra: 0.11; BKC: 0.10; Andheri: 0.12; Kurla: 0.09; Powai: 0.10; Ghatkopar: 0.09; Thane: 0.11; Kalyan: 0.10. These weights sum to 1.00 and are model inputs, not measured estimates of actual demand by Mumbai zone. Illustrative grid coordinates are CSMT (0,0), Dadar (1,1), Bandra (2,1), BKC (2,2), Andheri (3,2), Kurla (3,3), Powai (4,3), Ghatkopar (4,4), Thane (6,4), and Kalyan (8,5). They are not geographic coordinates or road distances.



International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Assumed hourly request profile

The baseline request counts for hours 00 through 23 are: 100, 70, 60, 50, 60, 100, 180, 250, 300, 280, 250, 260, 280, 300, 350, 400, 450, 500, 520, 480, 400, 320, 220, and 140. Their sum is 6,320 requests over the simulated 24-hour day. This hypothetical profile represents higher assumed activity in the afternoon and evening; it is not calibrated to observed Mumbai platform records.

Driver and request generation

The base scenario generates 500 drivers with assumed eight-hour shifts. Shift start hours are sampled from -7 to 23, allowing some shifts to overlap the start of the 24-hour window or extend beyond it. Initial driver locations are sampled according to zone weights. Each request receives a random timestamp within its hour, a pickup zone sampled using the same weights, and a destination sampled from the remaining zones so pickup and destination differ. Trip distance is the Euclidean distance between illustrative grid points multiplied by a random factor from 0.8 to 1.25. Trip duration is generated as $8 + 4 \times \text{distance units} + \text{a random value between 5 and 12 minutes}$. Synthetic driver pay is $\max(\text{₹}40, \text{₹}40 + \text{₹}15 \times \text{distance units} + \text{₹}2 \times \text{duration} + \text{random noise with standard deviation ₹}10)$.

Purpose and interpretation of the synthetic dataset

Generated fields include request ID, request time, pickup zone, destination zone, trip distance in illustrative units, trip duration in minutes, and synthetic driver pay in INR. These data are designed for comparative simulation, not for predicting actual fares or travel times. The random seed makes scenarios reproducible. Repeated experiments use seeds 10, 20, 30, 40, and 50; for each seed, baseline and fairness-aware allocation are tested against the same driver roster and request stream. Differences between strategies within a seed therefore reflect the allocation rule under that scenario, while differences across seeds reflect generated scenario variation.

Assumptions and limits of the data design

The synthetic data do not include observed Mumbai booking records, actual road-network routes, congestion, cancellations, platform commissions, tolls, tips, driver preferences, or validated market fare structures. Zone weights and the hourly profile are illustrative. Euclidean grid distance is only a proxy for spatial separation. The generated pay formula is a modelling choice, not an estimate of Uber, Ola, or another operator's pricing or driver compensation. Results must be described as outcomes under specified assumptions, not empirical facts about Mumbai drivers.

D. Allocation strategies

The baseline assigns each request to the nearest eligible available driver, where eligibility is determined by simulated shift status and current availability. The fairness-aware policy normalizes pickup distance and each eligible driver's cumulative simulated earnings, then minimizes a weighted score. In the tested 65/35 setting, distance receives weight 0.65 and cumulative earnings receives weight 0.35. Lower cumulative earnings can improve a driver's assignment priority, but the policy may select a driver who is farther from the pickup. The weight is an experimental parameter, not a learned or externally validated optimum.

Both strategies are evaluated on the same generated scenario for each random seed. The notebook uses five seeds: 10, 20, 30, 40, and 50. Outcomes include assigned requests, service rate, total assigned pay, mean and median simulated earnings per driver, earnings Gini, number of drivers with zero trips, mean trips per driver, mean pickup travel time, and the 90th percentile pickup time.

E. Fairness and efficiency metrics

Metric definitions. The Gini coefficient is $G = [\sum_i \sum_j |x_i - x_j|] / [2n \sum_i x_i]$, where x_i is a driver's cumulative simulated assigned pay and n is the number of simulated drivers. Service rate is the number of assigned requests divided by the total generated requests, multiplied by 100%. Mean pickup time is the arithmetic mean of simulated pickup-wait times for assigned requests. These metrics are computed separately; a lower Gini does not alone establish improved overall welfare.

The Gini coefficient is calculated over each simulated driver's cumulative assigned pay, including drivers with zero assigned pay. Lower Gini indicates a more equal distribution in the simulated scenario. Operational measures include assignment count, service rate, and pickup travel time. These metrics represent different objectives: a lower Gini does not by itself imply higher earnings for every driver or better overall system performance.



International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

F. Sensitivity experiments

The notebook compares four distance/earnings weight settings (90/10, 65/35, 50/50, and 30/70), varies the pickup-time multiplier (2, 3, and 4 minutes per illustrative distance unit), and explores low, normal, and high demand cases. The demand generator creates the high-demand case by adding sampled copies of existing requests; therefore, this is an exploratory stress test, not a realistic independent demand-generation model. It is retained as a limitation and future improvement.

IV. RESULTS & DISCUSSION

A. NYC trip-level payment patterns

In January 2025, the notebook reports 20,405,666 trips, with mean recorded driver pay of \$18.65 and median of \$13.94 across all records. The minimum is -\$55.02 and the maximum is \$994.20. There are 20,401,128 positive-pay trips. For the positive trip-pay distribution, the overall Gini coefficient is approximately 0.3969. This is a trip-level distribution measure and must not be interpreted as driver-level inequality.

Across the NYC records, the notebook reports a mean driver-pay value of \$18.65 and a median of \$13.94 per trip; the observed minimum is -\$55.02 and maximum is \$994.20. There are 20,401,128 positive-pay records. The Gini coefficient for positive trip-payment values is approximately 0.3969. This describes the spread of payments across trips and must not be presented as a measure of inequality among drivers.

Time band	Trips	Mean pay/trip positive	Median pay/trip positive	Trip-pay Gini
Early morning (00:00–05:59)	2,460,688	\$19.83	\$15.49	0.3878
Morning/daytime (06:00–11:59)	5,286,050	\$18.43	\$13.81	0.3929
Afternoon (12:00–16:59)	4,933,274	\$19.07	\$13.85	0.4077
Evening/night (17:00–23:59)	7,721,116	\$18.17	\$13.65	0.3943

The hourly summaries show that average positive payment per trip peaks at about \$21.44 at 5 AM and reaches a low of about \$17.28 at 7 PM. The largest hourly trip volume occurs at 6 PM, at roughly 1.26 million records. Hourly trip-payment Gini values range from about 0.3773 at 3 AM to 0.4143 at 3 PM. Of the four time bands defined in the notebook, afternoon has the highest trip-payment Gini (about 0.4077), while early morning has the highest mean positive payment per trip (about \$19.83). These are descriptive patterns; the analysis does not identify their causes.



International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

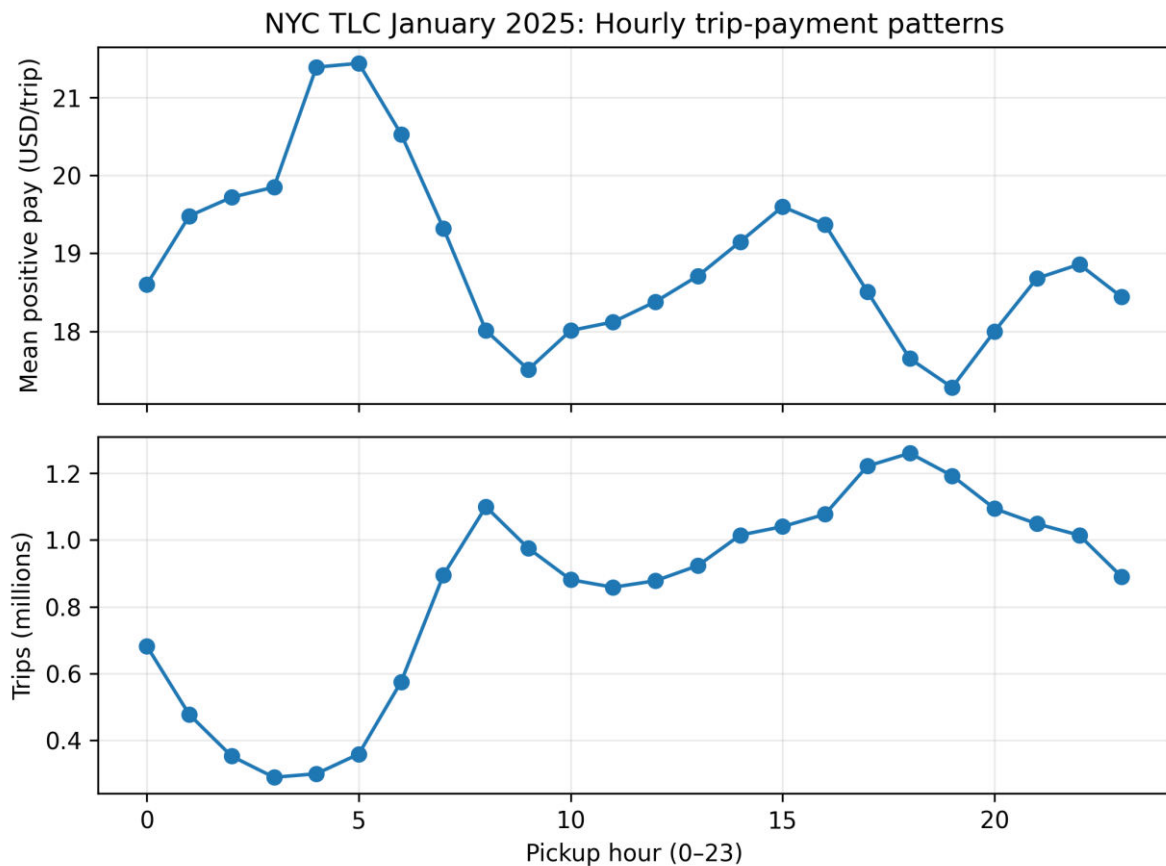


Fig. 1. Hourly mean positive driver payment per trip and trip volume in the January 2025 NYC TLC HVFHV records.

B. Baseline versus 65/35 fairness-aware policy

Metric (five-seed average)	Baseline	Fairness-aware 65/35
Assigned requests	3,999.4	3,970.6
Service rate	63.282%	62.826%
Earnings Gini	0.327	0.318
Drivers with zero trips	36.4	9.6
Mean pickup travel time	4.787 min	5.130 min
Total assigned pay	₹607,852.09	₹603,149.84

In the five-seed comparison, the fairness-aware rule lowered mean simulated driver-level earnings Gini from 0.327 to 0.318 and reduced the average number of drivers with no assigned trips from 36.4 to 9.6. It assigned about 28.8 fewer requests on average, corresponding to a service-rate decrease of approximately 0.456 percentage points. Mean pickup time increased by around 0.343 minutes, or 21 seconds. Gini was lower under the fairness-aware rule in each of the five tested seeds. This is evidence of consistency within the tested scenarios, not a formal significance test or proof that the result generalizes to other settings.

The change in allocation does not mean that all drivers earn more. The policy spreads assignments differently and reduces the number of drivers receiving none, but some summary measures—including median earnings in the initial single-seed comparison—can move downward. Results should therefore be discussed as a distributional change rather than a universal gain for every driver.



International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Five-seed average: nearest-driver baseline vs 65/35 policy

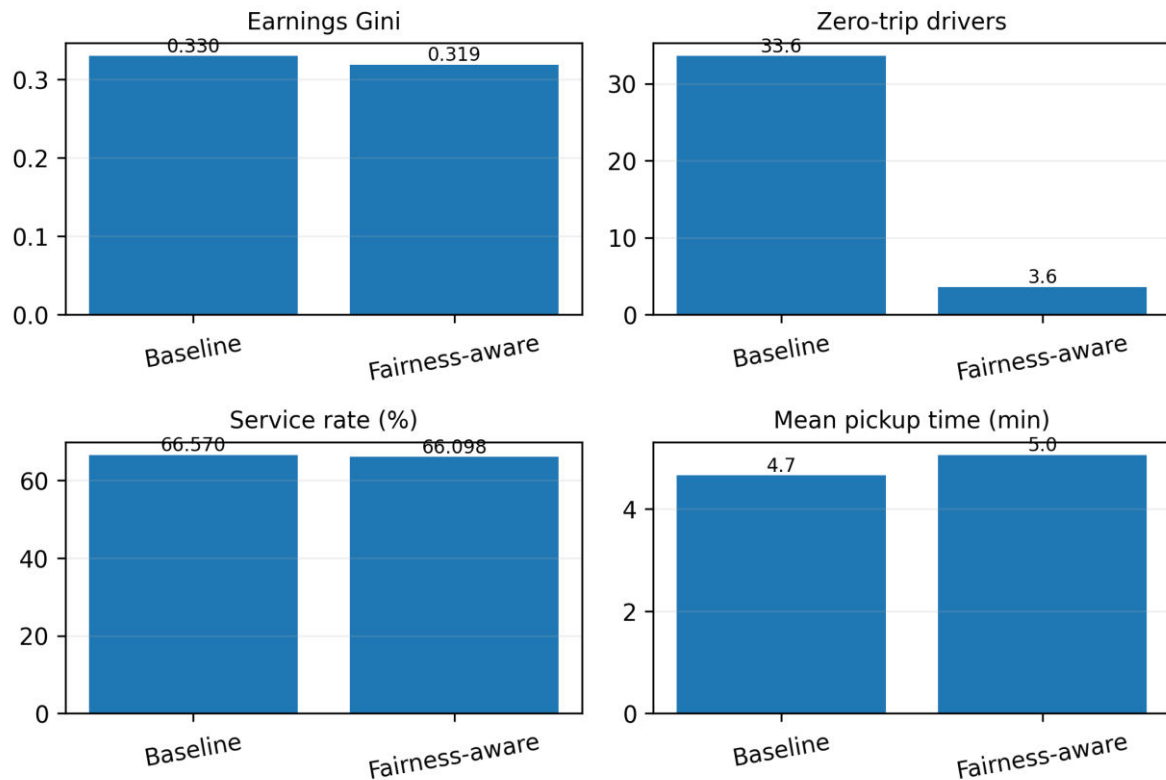


Fig. 2. Five-seed average comparison of the nearest-driver baseline and the 65/35 fairness-aware policy, using the corrected shift-boundary notebook outputs.

C. Fairness-weight experiment

Distance / earnings weight	Assigned requests	Service rate	Earnings Gini	Zero-trip drivers	Mean pickup time
90% / 10%	3,995.8	63.225%	0.318	10.8	4.775 min
65% / 35%	3,970.6	62.826%	0.318	9.6	5.130 min
50% / 50%	3,920.0	62.025%	0.315	9.2	5.856 min
30% / 70%	3,838.4	60.734%	0.312	9.0	6.911 min

Within the tested weight settings, assigning more importance to accumulated earnings tended to lower the simulated Gini while also reducing service rate and increasing pickup time. The 30/70 configuration yielded the lowest Gini among those tested, but also the weakest service rate and longest pickup time. The 65/35 configuration is retained as a practical compromise for comparison; these experiments do not establish it as an optimal setting.



International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

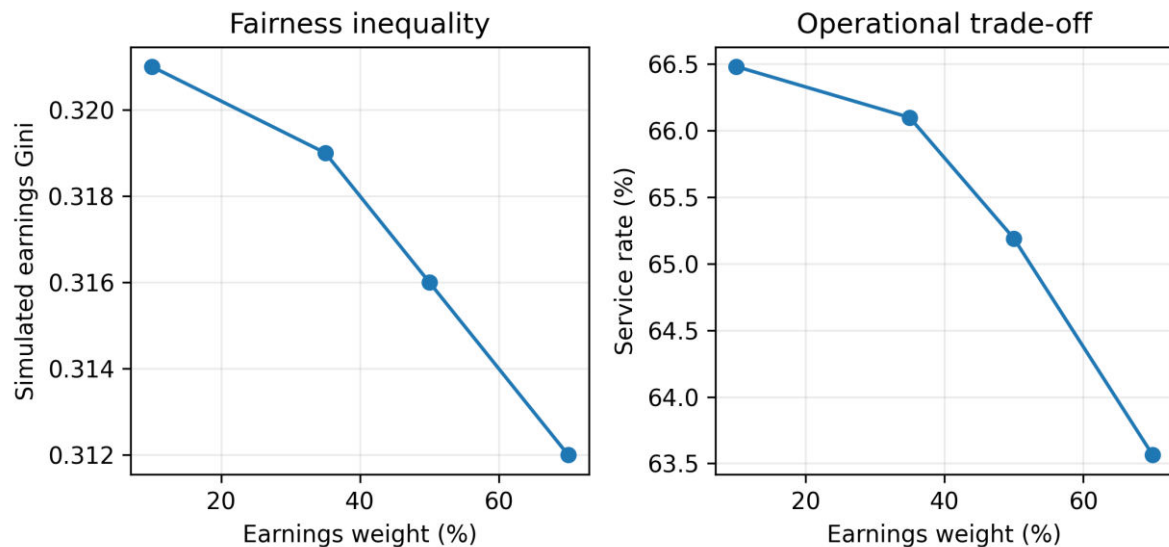


Fig. 3. Simulated earnings Gini and service rate across the tested earnings-weight settings.

D. Sensitivity analysis

When the pickup-time multiplier was varied from 2 to 4 minutes per illustrative distance unit, the fairness-aware policy retained a lower Gini than baseline in all three settings. Its mean pickup time was also higher than baseline in each setting. Since the multiplier is an assumed parameter and the coordinates are not road-network distances, this test establishes robustness only to the tested simulation parameter values.

Pickup-time factor	Baseline Gini	Fairness Gini	Baseline assigned	Fairness assigned	Fairness mean pickup
2.0	0.331	0.323	4,171.0	4,157.6	3.305 min
3.0	0.327	0.318	3,999.4	3,970.6	5.130 min
4.0	0.323	0.313	3,835.4	3,811.8	6.813 min

In the exploratory demand experiment, the fairness-aware policy had lower Gini and fewer zero-trip drivers in low-, normal-, and high-demand cases, with a small service-rate reduction. The high-demand case duplicates existing requests and produced 7,583 requests per seed rather than the intended 7,584 because of integer truncation. Therefore, the high-demand scenario and conclusions drawn from the demand experiment remain provisional; the count should be corrected and the experiment rerun before these results are treated as final.



International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

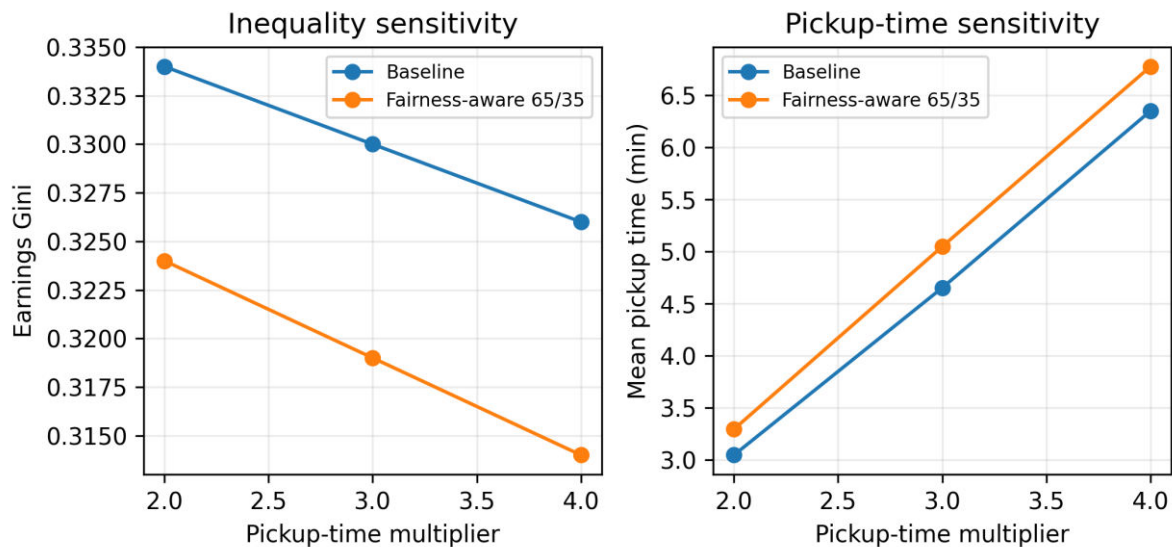


Fig. 4. Sensitivity of simulated earnings Gini and mean pickup time to the pickup-time multiplier, using the corrected notebook outputs.

Discussion of Findings

In the NYC descriptive analysis, trip payments differ by hour, time band, and pickup zone. Airport trips often have higher payment amounts per trip than shorter urban trips, but they also tend to involve greater distance and duration. Payment per trip alone is therefore an incomplete basis for comparing driver productivity or fairness. The available public records support trip-level description, not the estimation of driver-level earnings in this study.

The simulation indicates that considering cumulative earnings in dispatch can reduce concentration in simulated earnings and the number of drivers who receive no assignments. In the same scenarios, the policy has operational costs: average pickup time is longer and slightly fewer requests are served. These results suggest that dispatch fairness should be evaluated together with service outcomes. Before applying such a policy in practice, researchers would need credible online-time and earnings data, road-network travel estimates, realistic demand, operational constraints, and feedback from drivers and other stakeholders.

Study Limitations

- NYC TLC data are used as an external empirical benchmark, not as a representation of Mumbai.
- Trip-level payment Gini is not driver-level earnings Gini; the analysis does not estimate actual hourly driver earnings.
- The title's reference to Mumbai denotes a synthetic Mumbai-inspired scenario of ten representative zones, not a full-city geographic or road-network simulation. The Mumbai environment is synthetic; its zone weights, grid coordinates, demand curve, driver shifts, trip durations, and pay formula are assumptions.
- Euclidean grid distance does not represent Mumbai road-network distance or congestion.
- The pickup-time multiplier is assumed and is tested only over a limited range.
- The corrected notebook includes a shift-boundary eligibility check requiring estimated pickup wait plus trip duration to fit within the driver's shift. This code path should remain in the final reproducible notebook and be verified against the saved outputs.
- The high-demand scenario reuses sampled requests and is an exploratory stress test rather than an independently generated demand stream.
- Only five random seeds are used for the main repeated-scenario comparison; no formal inferential test is reported.
- The 65/35 weight is selected as a compromise among tested settings and is not established as globally optimal.
- The work does not establish causal discrimination, proprietary platform algorithmic bias, or real-world policy effectiveness.



International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

V. CONCLUSION

The paper pairs a descriptive study of January 2025 NYC TLC trip payments with a separate synthetic dispatch experiment inspired by Mumbai. The NYC analysis identifies variation in trip-level payments but does not establish driver-level earnings inequality. In the corrected simulation, the 65/35 distance/earnings policy reduced average simulated driver-level Gini from 0.327 to 0.318 and lowered the mean number of zero-trip drivers from 36.4 to 9.6. Service rate fell slightly and mean pickup time increased. The evidence supports further study of fairness-aware dispatch as a multi-objective problem; it is not evidence of a measured real-world impact.

Further work should use road-network travel times in place of grid distances, calibrate demand and pay assumptions against credible local evidence, improve the high-demand generator, and consistently enforce shift and trip-completion constraints. More random seeds and uncertainty estimates would strengthen the comparison, as would additional fairness measures and consultation with drivers. A structured review of the literature and final verification of bibliographic details are also needed before making a strong novelty claim or submitting the paper.

Data and reproducibility statement. The NYC TLC trip-record data are publicly available from the New York City Taxi and Limousine Commission at the URL listed in Reference [5]. The Mumbai-inspired scenario is synthetic and generated from the assumptions documented in this paper and the project notebook. The corrected notebook includes a shift-boundary eligibility check. Main comparison and pickup-time sensitivity outputs are reported from that notebook; the high-demand scenario remains provisional because its generator produces 7,583 requests rather than the intended 7,584 per seed.

REFERENCES

- [1] D. Shi, Y. Tong, Z. Zhou, B. Song, W. Lv, and Q. Yang, "Learning to assign: Towards fair task assignment in large-scale ride hailing," in Proc. 27th ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD '21), 2021, pp. 3549–3557, doi: 10.1145/3447548.3467085.
- [2] E. Bokányi and A. Hannák, "Understanding inequalities in ride-hailing services through simulations," Scientific Reports, vol. 10, Art. no. 6500, 2020, doi: 10.1038/s41598-020-63171-9.
- [3] J. Sun, H. Jin, Z. Yang, and L. Su, "Optimizing long-term efficiency and fairness in ride-hailing under budget constraint via joint order dispatching and driver repositioning," IEEE Transactions on Knowledge and Data Engineering, vol. 36, no. 7, pp. 3348–3362, Jul. 2024, doi: 10.1109/TKDE.2023.3348491.
- [4] M. Tzur and S. Ben-Gal, "Data-driven policies for the online ride-hailing problem with fairness," Transportation Science, published online Mar. 26, 2025, doi: 10.1287/trsc.2023.0068.
- [5] New York City Taxi and Limousine Commission, "TLC Trip Record Data." [Online]. Available: <https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page>. Accessed: Oct. 17, 2026.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com