Documentation of Rush-Hour Ridership Data Processing and methods to calculate total lane length in NYC

1 All results

Flow Type	Total	2024	2025
NN	17,391,444.37	8,606,270.672	8,785,173.702
CN	4,701,057.094	2,380,888.257	2,320,168.837
NC	17,423,562.76	8,734,065.838	8,689,496.923
CC	6,798,606.001	3,437,828.893	3,360,777.108

Table 1:

Flow type	All days (daily avg)	2024 (daily avg)	2025 (daily avg)
NN	404452.1947	391194.1215	418341.6048
CN	109326.9092	108222.1935	110484.2303
NC	405199.1340	397002.9926	413785.5677
CC	158107.1163	156264.9497	160037.0052

Table 2: Daily average ridership by flow type and year.

Total centerline road length in CRZ (441.66 km). The sum of all unique road centerline lengths within the CRZ, regardless of the number of lanes.

$$\text{CRZ total centraline length} \approx 441.66 \text{ kms} \approx 274.43 \text{ miles}.$$

Total lane-km of roads in CRZ (1019.67 km). The sum of "length $\times$ lane count" across all unique segments, representing the total available lane capacity.

$$\text{CRZ total road lane length} \approx 1019.67 \text{ kms} \approx 633.69 \text{ miles}.$$

Unique segments missing `lanes` tag (694). After deduplication, 694 physical segments have no lane information in OSM and are conservatively treated as having one lane.

2 Total counts of all MTA trips

This document describes the structure of the weekday rush-hour ridership dataset and the methodology used to construct four origin–destination flow groups and compute monthly ridership estimates for August 2024 and August 2025. The entire workflow is presented in a clear and systematic manner suitable for academic reporting or methodological documentation.

2.1 Structure of the Original Dataset

The dataset `df_rush_weekday` contains 6,372,016 observations and the following seven variables:

- `Timestamp`: character string.
- `ts`: hourly timestamp (`datetime64[us]`).
- `origin_nta`: origin NTA code.
- `destination_nta`: destination NTA code.
- `Estimated Average Ridership`: hourly ridership value.
- `is_origin_crz`: boolean indicator for whether the origin lies in the CRZ.
- `is_destination_crz`: boolean indicator for whether the destination lies in the CRZ.

The dataset includes:

- weekdays only,
- the 5:00–10:00 AM morning rush hour,
- observations from August 2024 and August 2025.

Important Property of the Time Variables

The columns `ts` and `Timestamp` do **not** contain all calendar days of each month. Instead, they include only:

> **the first complete weekday week (Monday–Friday) of each month**, with observations recorded at an hourly level.

Moreover, each row's `Estimated Average Ridership` value represents:

> the average hourly ridership for that month, computed across all days sharing the same weekday and hour.

Thus, the dataset contains "weekday-hour averages" rather than true daily ridership values.

2.2 Construction of Four Directional Subsets

Based on whether the origin and destination NTAs lie inside or outside the CRZ, the dataset is partitioned into four subsets:

- `nn_rush_weekday`: non-CRZ → non-CRZ,
- `cn_rush_weekday`: CRZ → non-CRZ,
- `nc_rush_weekday`: non-CRZ → CRZ,
- `cc_rush_weekday`: CRZ → CRZ.

Each subset preserves the same structure and variables as the original dataframe.

2.3 Expanding Weekly Observations to Monthly Estimates

Because the dataset contains observations only from each month's first complete weekday week, additional processing is required to estimate total weekday morning rush-hour ridership for the full month.

2.3.1 Step 1: Counting Weekday Occurrences Within Each Month

For every row in each subset, a new variable is added:

$$\texttt{weekday\_count\_in\_month},$$

which denotes the number of times the corresponding weekday occurs in that month.

For example, if a given month contains four Mondays, then all Monday rows in that month are assigned a value of 4.

2.3.2 Step 2: Computing Monthly Estimated Ridership

A second new variable is added:

$$\texttt{EAR\_MONTH} = \texttt{Estimated Average Ridership} \times \texttt{weekday\_count\_in\_month},$$

representing the estimated monthly total ridership for that weekday–hour combination.

After this computation, four enhanced dataframes are obtained:

- `noncrz_to_noncrz_rush_weekday`,
- `crz_to_noncrz_rush_weekday`,
- `noncrz_to_crz_rush_weekday`,
- `crz_to_crz_rush_weekday`,

each containing the additional variables `weekday_count_in_month` and `EAR_MONTH`.

2.4 Monthly Ridership Computation for August 2024 and 2025

For each of the four directional flow groups, three aggregated quantities are computed:

1. the total weekday (5–10 AM) estimated ridership in August 2024;
2. the total weekday (5–10 AM) estimated ridership in August 2025;
3. the combined total ridership for both years.

This yields twelve summary statistics in total:

Non-CRZ → Non-CRZ (NN)

`total_counts_nn`, `total_counts_nn_2024`, `total_counts_nn_2025`.

CRZ → Non-CRZ (CN)

`total_counts_cn`, `total_counts_cn_2024`, `total_counts_cn_2025`.

Non-CRZ → CRZ (NC)

`total_counts_nc`, `total_counts_nc_2024`, `total_counts_nc_2025`.

CRZ → CRZ (CC)

`total_counts_cc`, `total_counts_cc_2024`, `total_counts_cc_2025`.

Together, these twelve summary statistics characterize weekday morning rush-hour mobility flows between CRZ and non-CRZ regions for August 2024 and August 2025. And

Flow Type	**Total**	**2024**	**2025**
NN	17,391,444.37	8,606,270.672	8,785,173.702
CN	4,701,057.094	2,380,888.257	2,320,168.837
NC	17,423,562.76	8,734,065.838	8,689,496.923
CC	6,798,606.001	3,437,828.893	3,360,777.108

Table 3:

divided each number above by the respected number of weekdays gives the average daily data.

Flow type	All days (daily avg)	2024 (daily avg)	2025 (daily avg)
NN	404452.1947	391194.1215	418341.6048
CN	109326.9092	108222.1935	110484.2303
NC	405199.1340	397002.9926	413785.5677
CC	158107.1163	156264.9497	160037.0052

Table 4: Daily average ridership by flow type and year.

2.5 Verification

2.5.1 method 1

We have verified that the official 2024 *average weekday ridership* reported by the MTA, denoted by $R_{\text{MTA}} = 3{,}735{,}571$, is very close to the corresponding estimate produced by our

method, $\hat{R} = 3{,}706{,}754$. The relative error is

$$\text{RE} = \frac{|R_{\text{MTA}} - \hat{R}|}{R_{\text{MTA}}} = \frac{|3{,}735{,}571 - 3{,}706{,}754|}{3{,}735{,}571} \approx 0.0077,$$

i.e., about 0.77%. This comparison is based on weekday ridership over all months of 2024 and across the full 24-hour period.

2.5.2 method 2(Cross-Check Based on Daily Series)

To assess the reliability of the above method, I perform an independent validation directly on the unaggregated raw data, without relying on month-level totals:

1. **Filtering the raw data**
   Starting from the original OD dataset, I filter observations to:
   - years 2024 and 2025,
   - month = August,
   - weekdays only,
   - time window 5–10am.
2. **Grouping by OD flow type**
   I then partition the filtered data into four OD flow categories:
   - NN: non-CRZ → non-CRZ,
   - CN: CRZ → non-CRZ,
   - NC: non-CRZ → CRZ,
   - CC: CRZ → CRZ.
3. **Daily aggregation**
   For each flow category, I aggregate by calendar date: for a given date, I sum `Estimated Average Ridership` across all 5–10am time buckets, obtaining the *total* estimated ridership for that date and flow type.
4. **August daily averages**
   Finally, for each flow category, I take the simple arithmetic mean of these daily totals over all weekdays in August (2024 and 2025). This yields the quantity *"Mean daily summed Estimated Average Ridership in August"* for each of the four flows:

- CN (CRZ → non-CRZ): 109,720.56194,
- NC (non-CRZ → CRZ): 408,530.42970,
- CC (CRZ → CRZ): 160,035.77351,
- NN (non-CRZ → non-CRZ): 405,189.68832.

5. **Comparison with the original method**
   The original, month-based method (which reconstructs monthly totals using

   $$\text{average}(\text{year}, \text{month}, \text{weekday}, \text{time-of-day}) \times \text{number of occurrences of that weekday in the month},$$

   followed by division by the total number of weekdays) produces the following daily averages for August:

   - NN: 404,452.19474,
   - CN: 109,326.90917,
   - NC: 405,199.13397,
   - CC: 158,107.11631.

   These values are close to the corresponding daily means obtained from the validation method in Section 2.

6. **Methodological differences**
   The original method explicitly incorporates the number of times each weekday occurs in a given month by multiplying the weekday-level average by the actual weekday count and then normalising by the total number of weekdays. In contrast, the validation method works directly with the daily series, first constructing daily totals and then taking a simple average across all observed weekdays in August. Although the two approaches differ in how they construct monthly totals, they are mathematically consistent in how they aggregate over actual dates.

7. **Conclusion**
   Given that the two independently derived sets of daily averages are highly consistent across all four flow categories, I conclude that the original computation method used in Question 1 provides an accurate and trustworthy estimate of the daily average `Estimated Average Ridership` for the specified time window.

3 Calibration of Lane Counts Using Official NYC Statistics and OSM Data

3.1 Part 1: Official Baseline and Initial OSM Estimates

As a starting point, I take the official New York City roadway statistics as a ground-truth baseline. According to published figures, the total lane length of roads in New York City is approximately

$$19{,}000 \text{ lane-miles} \approx 30{,}571 \text{ lane-km},$$

while the total centerline road length is about

$$6{,}300 \text{ miles} \approx 10{,}136.7 \text{ km}.$$

From these two numbers, the length-weighted average number of lanes per road segment in NYC can be approximated by

$$\bar{\ell}_{\text{NYC}} \approx \frac{19{,}000}{6{,}300} \approx 3.02.$$

These values are regarded as authoritative and serve as the benchmark in the subsequent calibration.

Using OpenStreetMap (OSM) data with `network_type = "drive"` to estimate the NYC road network, the extracted centerline road length is

$$L_{\text{centerline}}^{\text{OSM}} \approx 11{,}130.73 \text{ km},$$

which is quite close in magnitude to the official 10,136.7 km. This suggests that OSM covers almost all actual road segments reasonably well in terms of geometry.

However, when the same OSM data are used to compute total lane length by directly multiplying each segment length by its recorded lane count, the result is only

$$L_{\text{lane}}^{\text{OSM}} \approx 16{,}193.15 \text{ lane-km},$$

which is far below the benchmark 30,571 lane-km. Given that the centerline length is close to the official value, the most plausible explanation is not missing road geometry, but rather that many segments have no lane information recorded in OSM.

3.2 Part 2: Structure of Missing Lane Tags (NYC vs. CRZ)

NYC-wide Missingness Pattern

Within the NYC-wide OSM road network:

- Number of road segments *without* any lane tag: 60,215.
- Number of road segments *with* a lane tag: 32,751.

Table 5 summarizes, by `highway_norm` class, how many segments are missing lane information, while Table 6 shows, for segments with lane tags, the number of segments and the mean lane count.

Table 5: NYC segments without lane tags by `highway_norm`.

`highway_norm`	Count without lane
residential	47,921
secondary	3,294
primary	1,436
tertiary	4,644
motorway_link	222
motorway	1
unclassified	807
trunk	10
primary_link	1,436
secondary_link	172
tertiary_link	127
trunk_link	53
living_street	92
busway	0

The majority of missing lane tags are concentrated in `residential` streets, while high-class roads (`primary`, `secondary`, `trunk`, `motorway`) have relatively better lane coverage and larger mean lane counts. The overall average lane count among observed segments is about 2.164 in NYC.

CRZ Subregion

A similar analysis in the Congestion Relief Zone (CRZ) yields:

- Number of segments *without* lane tags in CRZ: 694.
- Number of segments *with* lane tags in CRZ: 3,206.

Table 6: NYC segments with lane tags: counts and mean lanes.

highway_norm	Count with lane	Mean lanes
residential	10,176	1.49990173
secondary	7,709	2.567259048
primary	5,789	3.119018829
tertiary	4,726	1.926999577
motorway_link	1,529	1.232831916
motorway	1,124	2.972419929
unclassified	737	1.8385346
trunk	437	3.274599542
primary_link	232	1.172413793
secondary_link	113	1.150442478
tertiary_link	94	1.29787234
trunk_link	55	1.454545455
living_street	23	1.0
busway	7	1.0

Tables 7 and 8 show the corresponding breakdown.

Again, missing lane information is mainly found on `residential` streets, and the pattern of mean lane counts by road class in CRZ closely resembles the NYC-wide pattern.

3.3 Part 3: Six Imputation and Calibration Schemes

To correct for missing lane information in OSM, I consider six different imputation schemes for assigning a lane value x to segments without lane tags.

1. **Method 1: Simple constant imputation.**
   All segments without lane tags are assigned the constant value

   $$x = 3.02,$$

   i.e. the official citywide average lanes per road.

2. **Method 2: Constant x chosen so that the overall mean equals 3.02.**
   Let $N_{\text{miss}} = 60{,}215$ and $N_{\text{obs}} = 32{,}751$ denote the numbers of segments without and with lane tags in NYC, respectively. Let $\bar{\ell}_{\text{obs}}$ be the overall average lane count among observed segments. The constant x for all missing segments is chosen to satisfy

   $$x \cdot \frac{N_{\text{miss}}}{N_{\text{miss}} + N_{\text{obs}}} + \bar{\ell}_{\text{obs}} \cdot \frac{N_{\text{obs}}}{N_{\text{miss}} + N_{\text{obs}}} = 3.02.$$

Table 7: CRZ segments without lane tags by `highway_norm`.

`highway_norm`	Count without lane
residential	562
secondary	49
primary	31
trunk	1
motorway_link	7
unclassified	24
tertiary	15
motorway	0
living_street	3
trunk_link	0
primary_link	1
secondary_link	1
tertiary_link	0

3. **Method 3: Use the observed overall mean.**
   Every missing segment is assigned

$$x = \bar{\ell}_{\text{obs}},$$

   which implicitly assumes that segments with lane tags are representative of all segments.

4. **Method 4: Type-specific mean imputation.**
   For each `highway_norm` class h, compute the mean lane count $\bar{\ell}_h$ among segments of that class with lane tags. For a missing segment e with type h, set

$$x_e = \bar{\ell}_h.$$

   This preserves differences across road classes but does not perform any global scaling.

5. **Method 5: Scaled type-specific mean so that the *missing subset* has mean 3.02.**
   Start from Method 4, then multiply all type-specific means used for missing segments by a common factor so that, when considering only the missing subset, the average lane count becomes exactly 3.02. In other words,

$$x_e = \lambda \cdot \bar{\ell}_{\text{type}(e)}, \quad e \in \text{missing subset},$$

Table 8: CRZ segments with lane tags: counts and mean lanes.

highway_norm	Count with lane	Mean lanes
residential	1,092	1.305860806
secondary	816	2.743872549
primary	769	3.169050715
trunk	170	3.594117647
motorway_link	97	1.412371134
unclassified	93	1.494623656
tertiary	68	1.970588235
motorway	55	2.472727273
living_street	18	1.0
trunk_link	16	1.4375
primary_link	5	1.4
secondary_link	5	1.4
tertiary_link	2	1.0

where λ is chosen such that the mean of x_e over all missing segments equals 3.02. This retains cross-class structure and aligns the missing subset with the official citywide average.

6. **Method 6: Scaled type-specific mean so that the *full sample* has mean 3.02.** Again start from Method 4, but choose a scaling factor λ so that the combined sample (observed + imputed) has an overall mean of 3.02 lanes. Thus,

$$x_e = \lambda \cdot \overline{\ell}_{\text{type}(e)}$$

for missing segments e, with λ determined by the full-sample mean constraint.

When applied to the full NYC network, the average lane count among observed segments is about 2.164, and in the CRZ it is about 2.281. These are close, suggesting that the observed subset has a similar lane structure in both areas. Under the assumption that the true citywide mean is ≈ 3.02, using 3.02 as a calibration target is reasonable; in particular, it is also reasonable to assume that the "true" average lane count in the CRZ is close to 3.02, provided that missing-lane roads are distributed over road classes in a similar way as in NYC as a whole.

3.4 Part 4: Numerical Results and Preferred Scheme

For each method, I compute:

- The estimated total lane-kilometers in NYC and its relative error compared to the benchmark 30,571 lane-km.
- The estimated total lane-kilometers in the CRZ when the same method is applied but with the network restricted to the CRZ.

The results are summarized in following:

Table 9: Comparison of different methods

Method	Value	Error	Rank
method_1	29904.05	0.021816427	1
method_2	33194.67	0.085822184	4
method_3	23979.62	0.215608910	5
method_4	20366.25	0.333804913	6
method_5	29547.84	0.033468320	2
method_6	32780.65	0.072279284	3

Interpretation.

- In terms of NYC-wide accuracy, **Method 1** (constant $x = 3.02$ for all missing segments) produces the smallest relative error ($\approx 2.1\%$) with respect to the benchmark 30,571 lane-km.
- **Method 5** also yields a very small error ($\approx 3.3\%$), clearly outperforming Methods 2, 3, 4, and 6.
- However, Method 1 ignores all heterogeneity across road classes by forcing every missing segment to have exactly 3.02 lanes, which reduces interpretability.
- Method 5, in contrast, preserves class-specific lane patterns via type-specific means and applies a common scaling only to the missing subset so that their average aligns with the official citywide mean of 3.02 lanes. It thus combines good numerical accuracy with a more realistic representation of the road hierarchy.

Conclusion. Although Method 1 achieves the smallest numerical error when validated against the NYC benchmark, Method 5 offers a better balance between accuracy and structural plausibility. It explicitly respects differences between `highway_norm` classes while still calibrating the missing subset to the official global mean. Therefore, **Method 5 is chosen as the preferred imputation strategy**, especially for extrapolating to the CRZ.

Under Method 5, the estimated total lane length in the CRZ is

$$\text{CRZ total road lane length} \approx 1019.67 \text{ kms} \approx 633.69 \text{ miles}.$$

This value is consistent with the NYC benchmark after calibration and preserves the observed OSM structure of lane counts across road types within the CRZ.
Also,

$$\text{CRZ total centraline length} \approx 441.66 \text{ kms} \approx 274.43 \text{ miles}.$$

4 Results Visualization

Figure 1: Figure 1 presents a bar chart of total flows across the four OD categories—CRZ→CRZ, CRZ→nonCRZ, nonCRZ→CRZ, and nonCRZ→nonCRZ—during weekday morning peak hours (5–10 AM) in August 2024 and August 2025.

Figure 2: Figure 2 shows the corresponding share distribution (percentage contribution) of each OD category for the same period.

Figure 3: Figure 3 separates the two years and displays each year's flow shares in two individual pie charts, one for 2024 and one for 2025.

Figure 4: Figure 4 provides a side-by-side bar chart comparing the OD flows between August 2024 and August 2025.

Figure 5: Figure 5 is the set of figure 1,2,3,4 but daily average not monthly sum, by simply divided by the number of weekdays.

Figure 6: Figure 6 presents the weekday morning-peak (5–10 AM) time series from January to August of 2024 and 2025. Note that the time axis is not continuous, because each month includes only data from its first complete week. For example: (1) the plot jumps directly from the first Friday of January 2024 to the first Monday of the first complete week of February 2024; (2) similarly, it jumps from the first complete week of August 2024 directly to the first complete week of January 2025.

The y-axis values are constructed as follows (illustrated here using the nonCRZ→nonCRZ rush-hour weekday flows, but applied to all months, weekdays, and 5–10 AM records). For each month and each weekday, all 5–10 AM observations of a given OD category are summed

(using the EAR_MONTH values) to obtain the total flow for that weekday. The date used on the horizontal axis corresponds to the actual weekday date within the month's first complete week.

Figure 7: Figure 7 shows the road network within the CRZ zone. Road segments with colors closer to yellow represent higher lane density.

Figure 8: Figure 8 presents The spatial distribution of every subway station in New York City on the map.

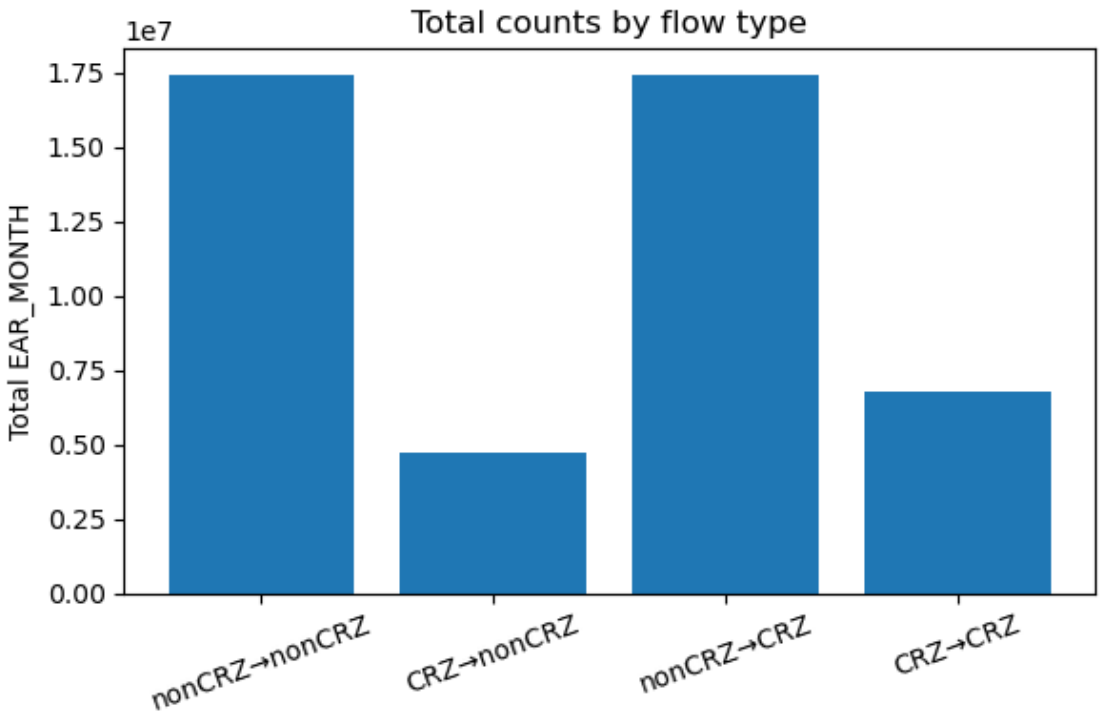

Figure 1:

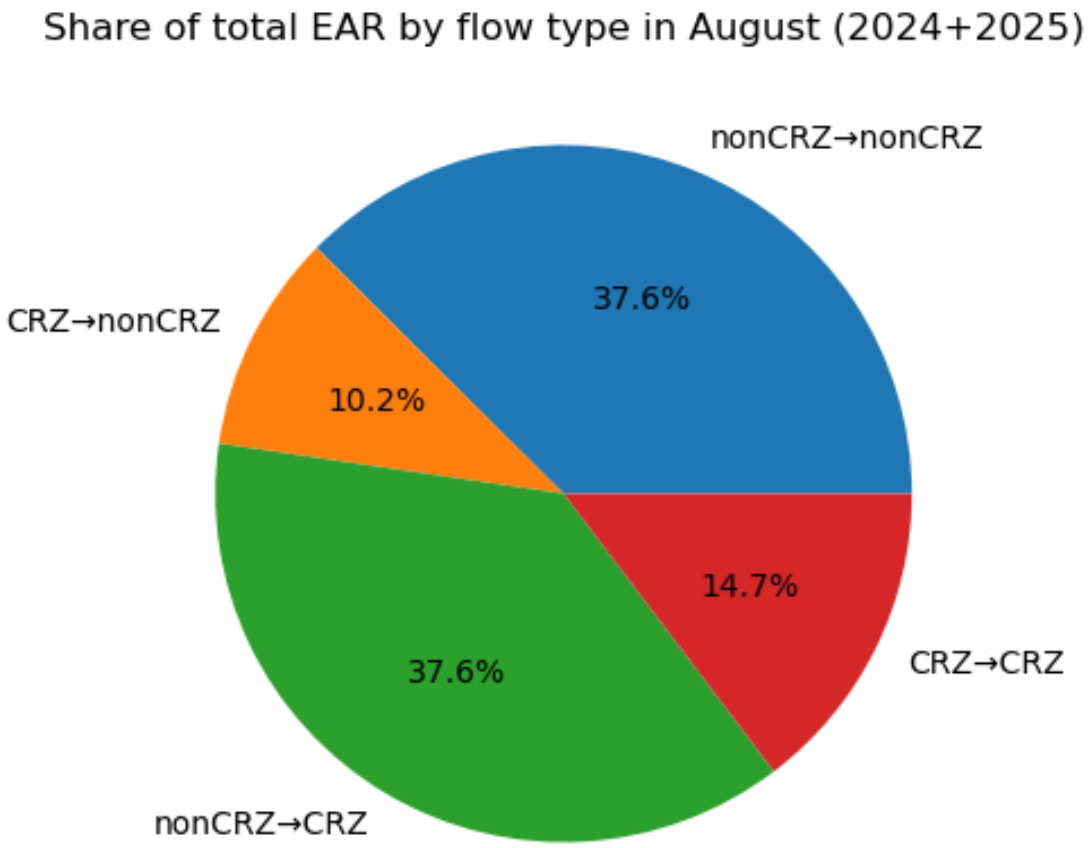

Figure 2:

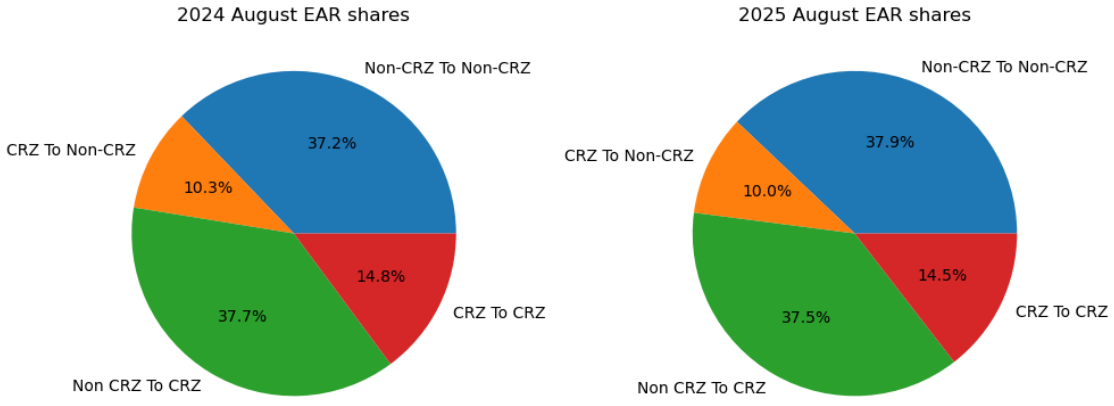

Figure 3:

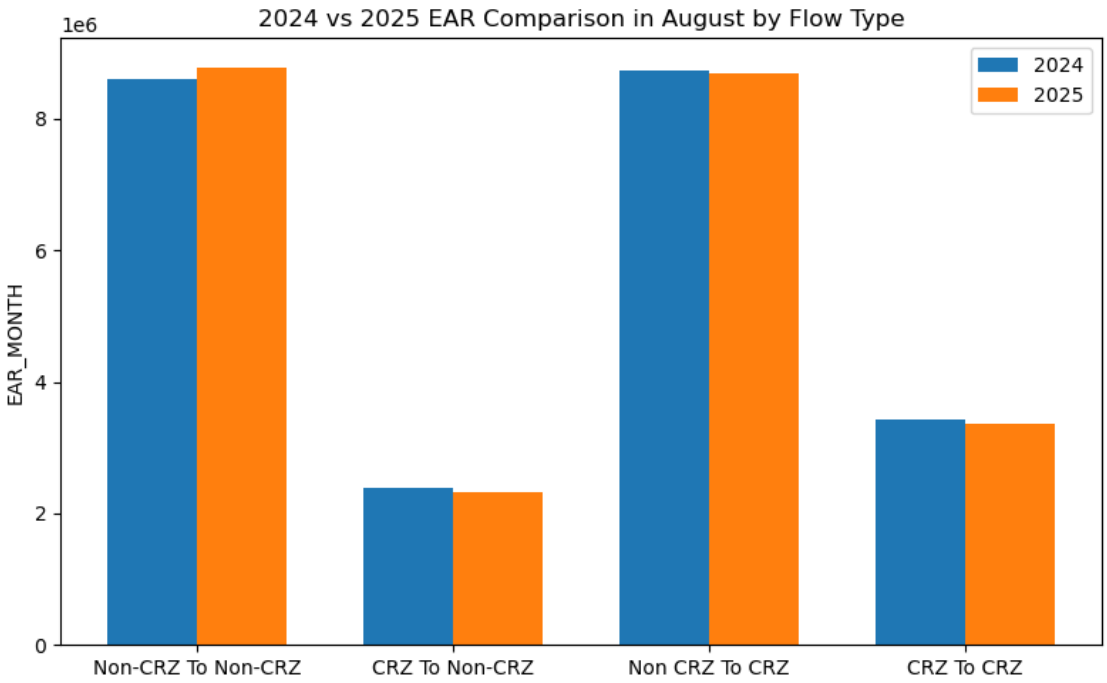

Figure 4:

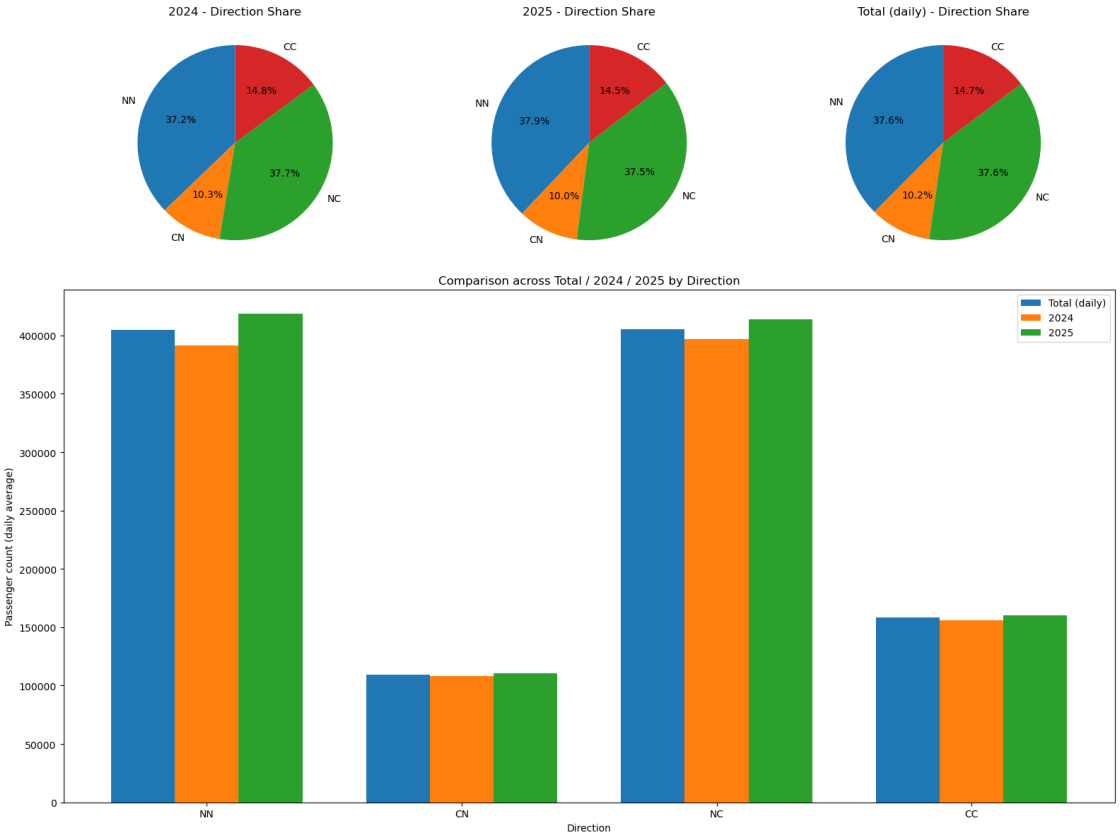

Figure 5:

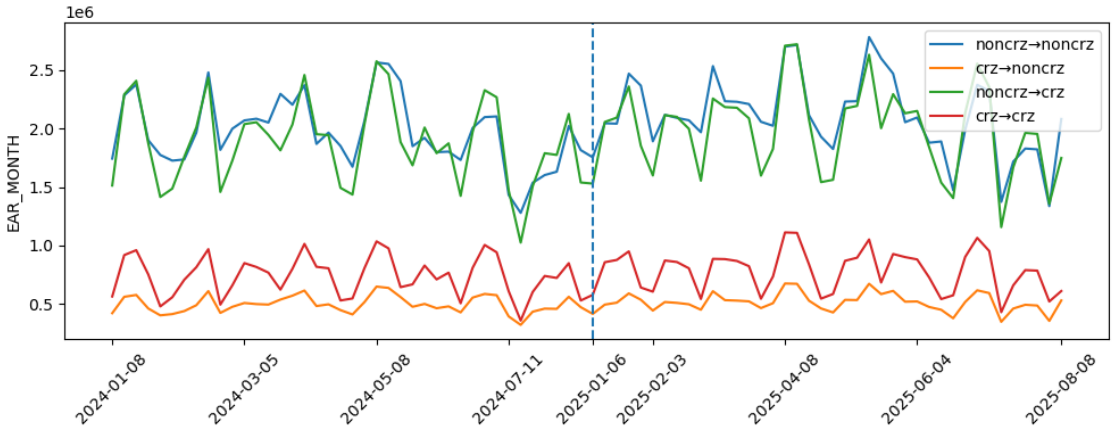
1e6
2.5
2.0
1.5
1.0
0.5
EAR_MONTH
noncrz→noncrz
crz→noncrz
noncrz→crz
crz→crz
2024-01-08
2024-03-05
2024-05-08
2024-07-11
2025-01-06
2025-02-03
2025-04-08
2025-06-04
2025-08-08

Figure 6:

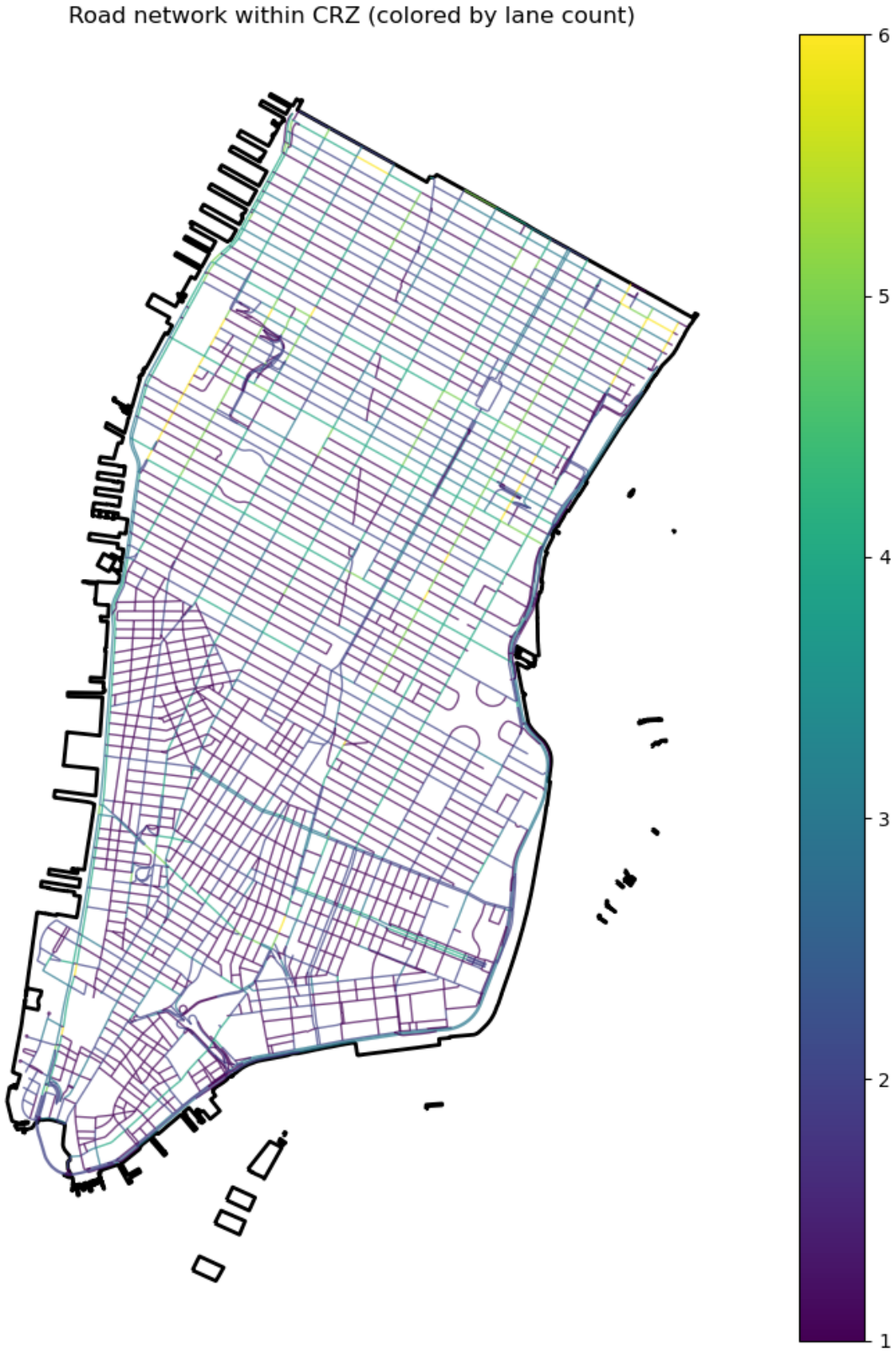

Figure 7:

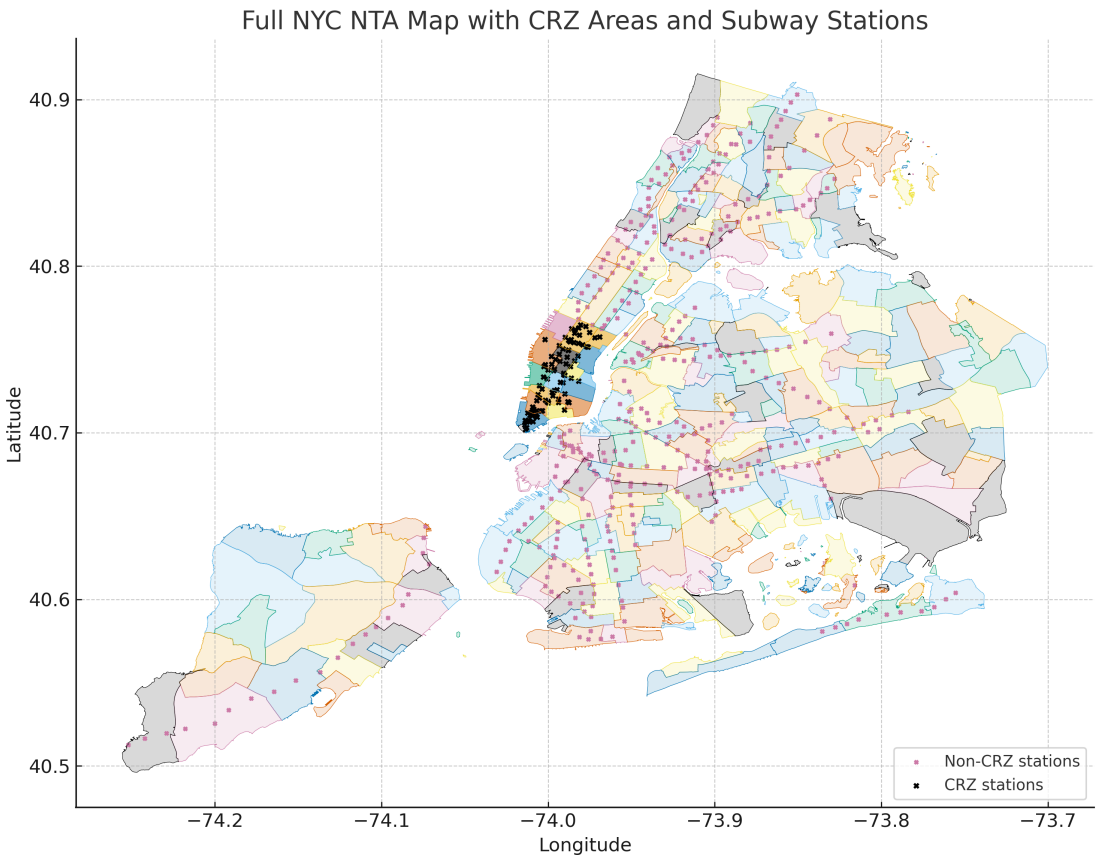
Full NYC NTA Map with CRZ Areas and Subway Stations
Latitude
Longitude
40.9
40.8
40.7
40.6
40.5
−74.2
−74.1
−74.0
−73.9
−73.8
−73.7
Non-CRZ stations
CRZ stations

Figure 8: