SoCal Transportation Study: Sampling, Data Processing, and Preliminary Results

Waves 1-3 (2024-2026)

RSG
the science of insight®

SCAG Modeling Task Force

April 23, 2026

RSG Team Introductions

Joe Chestnut, PhD
Senior Analyst
Travel Behavior Research

Joe Trost
Analyst
Travel Behavior Research

Youme Yai
Senior Consultant
Travel Behavior Research

RSG

Agenda

SCAG | SoCal Transportation Study

Metro Partnering with your community

Introduction	5 minutes
Sampling Plan Overview	5 minutes
Quality Assurance/Control Procedures	8 minutes
Weighting Methodology	14 minutes
Initial Findings from the Study	8 minutes
Q&A	5 minutes

RSG

Sampling Plan Overview

Address-Based Sampling

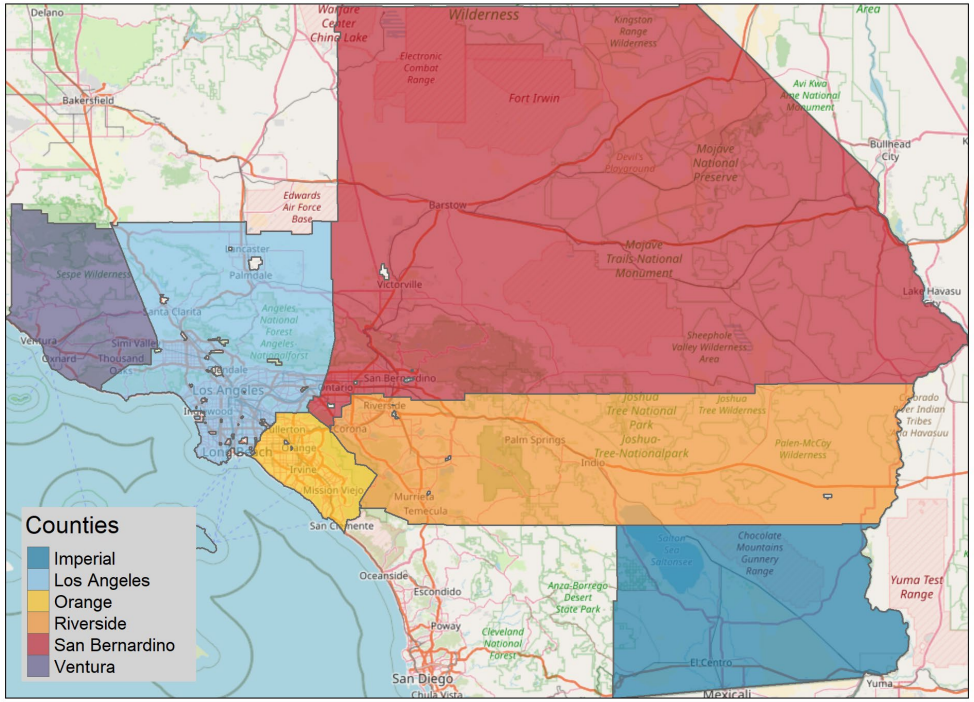
RSG uses address-based sampling (ABS) to collect a random sample of households.

ABS is used by drawing a random sample of addresses from all residential addresses in the survey region.

The size and distribution of this sample is determined by anticipated response rates and the distribution of ‘hard-to-survey’ households.

The selected addresses are mailed an invitation to the study followed by a reminder 1 week later.

All households within each Census block group have an equal chance of selection for sampling. Group quarters are excluded.



Stratification Variables

RSG uses demographics to identify hard-to-survey areas.

These areas are assumed to have lower response rates and are therefore sent more invitations.

For SCAG they were defined as block groups where:

- at least 20% of households earning less than the 2024 federal poverty level, or
- at least 20% of households having zero vehicles

Sending more invitations to these areas helps to promote more complete households and better representativeness.

Household Size	Household Income
1	$15,060
2	$20,440
3	$25,820
4	$31,200
5	$36,580
6	$41,960
7	$47,340
8	$52,720

PRESENTED BY JOE CHESTNUT

Recruitment Methods and Response Monitoring

RSG SoCal Transportation Study Wave 3 | Overview | Recruit/Conversion/Response Rates | Demographics

3,750
Survey Target Completes

2,545
Completes Total

275
Wave 3 Target Completes

872
Recruits To-Date

384
Completes To-Date

The following results exclude all test passwords and are pre-data cleaning.
Data were last updated on 2026-04-17

Study Name Results

Response Summary To-Date | Invitation Schedule | Survey Participation

Mail Group	Invitations	Recruits (Not Expired)	Recruits (Expired)	Recruit Rate	Completes (Not Expired)	Completes (Expired)	Completion Rate	Conversion Rate	Rmove Recruit Share	Online Recruit Share	Rmove Complete Share	Online Complete Share
1	8,194	0	205	2.5%	0	92	1.1%	43.9%	20.5%	79.5%	29.3%	70.7%
2	11,474	0	306	2.7%	0	146	1.2%	46.1%	14.7%	85.3%	16.4%	83.6%
3	13,112	0	311	2.4%	0	133	1.0%	41.5%	15.4%	84.6%	17.3%	82.7%
Email 1	300	0	11	3.7%	0	1	0.3%	9.1%	9.1%	90.9%	0.0%	100.0%
Email 2	600	0	30	5.0%	0	9	1.5%	30.0%	43.3%	56.7%	22.2%	77.8%
Email 3	600	0	9	1.5%	0	3	0.5%	33.3%	55.6%	44.4%	66.7%	33.3%
Total	34,280	0	872	2.5%	0	384	1.1%	42.8%	17.7%	82.3%	20.3%	79.7%

Note:

Invitations: Number of households scheduled to receive invitations
Recruits Not Expired: Number of recruited households whose travel diaries have not expired (7 days after travel period ended)
Recruits Expired: Number of recruited households whose travel diaries have expired
Recruit Rate: % of invited households that have recruited to-date (total). These will continue to rise throughout the study.
Completes Not Expired: Number of completed households whose travel diaries have not expired
Completes Expired: Number of completed households whose travel diaries have expired
Reinvites: Number of reinvited households
Reinvites Completed: Number of reinvited households that completed the survey
Conversion Rate: % of recruited households with expired travel periods that completed the study, NOT including reinvited households ((Completes Expired - Reinvited Completes)/ Recruits Expired)
Completion Rate: % of invited households who have completed the study NOT including reinvited households (Recruit Rate * Conversion Rate)
Reinvite Completion Rate: % of reinvited households who have completed the study (Reinvites Completed / Reinvites)

Key Challenges & Mitigation Strategies

Survey response was a challenge, particularly in Spanish speaking areas.

Initial survey waves indicated that survey responses were lower than anticipated.

This was particularly pronounced in Spanish speaking areas.

To address this RSG did the following:

- Redesigned the invitation materials to focus on graphics not photos
- Removed the invitation letter and replaced it with a postcard
- Designed Spanish language postcards for Imperial county

Before

After

Summary of Achieved Sample vs Targets

COUNTY	COUNTS	TARGET COMPLETES	PERCENT OF TARGET
Imperial	9	21	42.86%
Los Angeles	1,437	1,472	97.62%
Orange	475	469	101.28%
Riverside	273	332	82.23%
San Bernardino	251	290	86.55%
Ventura	109	121	90.08%
Total	2,554	2,705	94.42%

Quality Assurance & Control Procedures

Apply Real-Time Validation During Data Collection

Real-Time Logic Validation	Geocoded Location Validation	Trip-Level Logic Checks	Continuous Monitoring and QA
Survey tools validate inputs instantly ensuring data is **complete, consistent**, and **reasonable** at collection.	Locations such as home, work, and trip points are **geocoded and verified** during entry to ensure accuracy.	Checks prevent **overlapping** trips, **negative** durations, and **missing** timestamps for reliable trip data.	Live data monitoring and pattern analysis **detect anomalies early**, enabling timely corrective actions.

Initial Post-Field Data Review

Household certification and geographic validation

Household Eligibility Review

Households are certified based on completion of travel data and verified home location within the study region.

Data Quality and Privacy

Records with duplication, technical issues, or privacy deletion requests are removed to maintain data integrity.

Geographic Data Assignment

Geographic variables are assigned by spatially joining coordinates to census and PUMA boundaries using reported home locations.

Automated QA/QC and Data Pipeline

Machine-assisted review and scripted checks

Machine Learning Trip Review

Identifies trips needing manual review for anomalies like false stops, merging, or GPS drift.

Visual Reviews

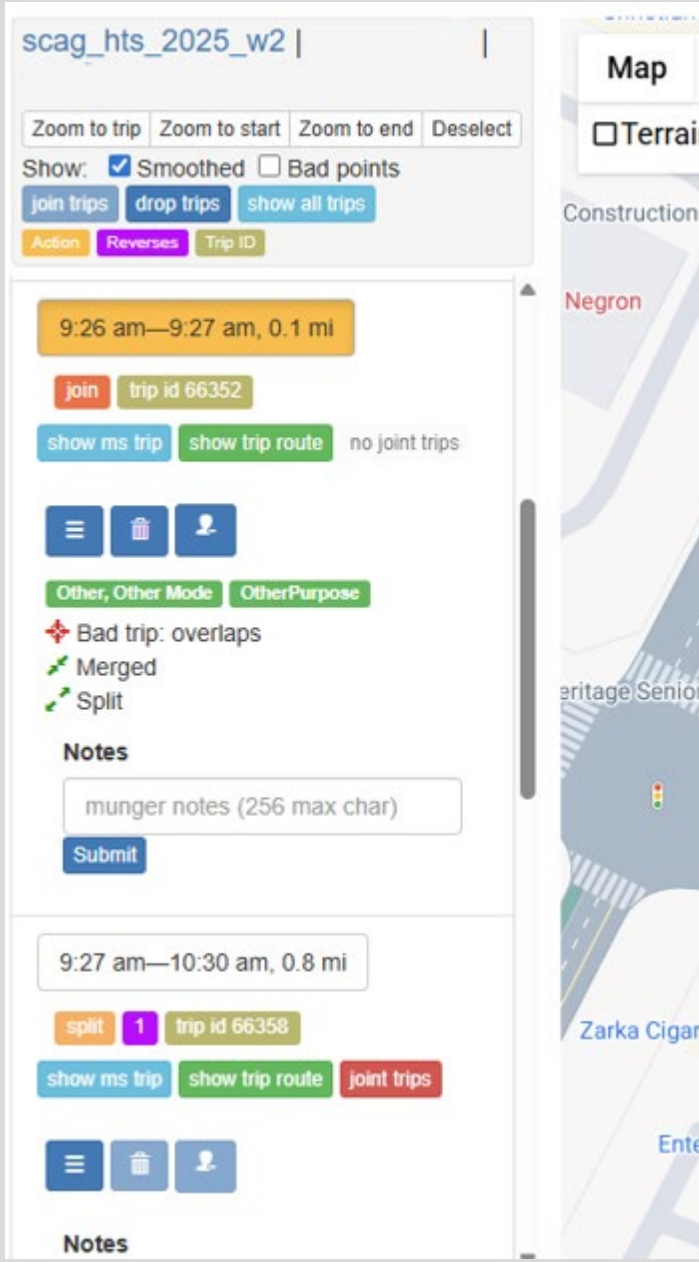
Flagged trips are overlayed on maps to visually check trip segments, paths, and times all appear to be correct.

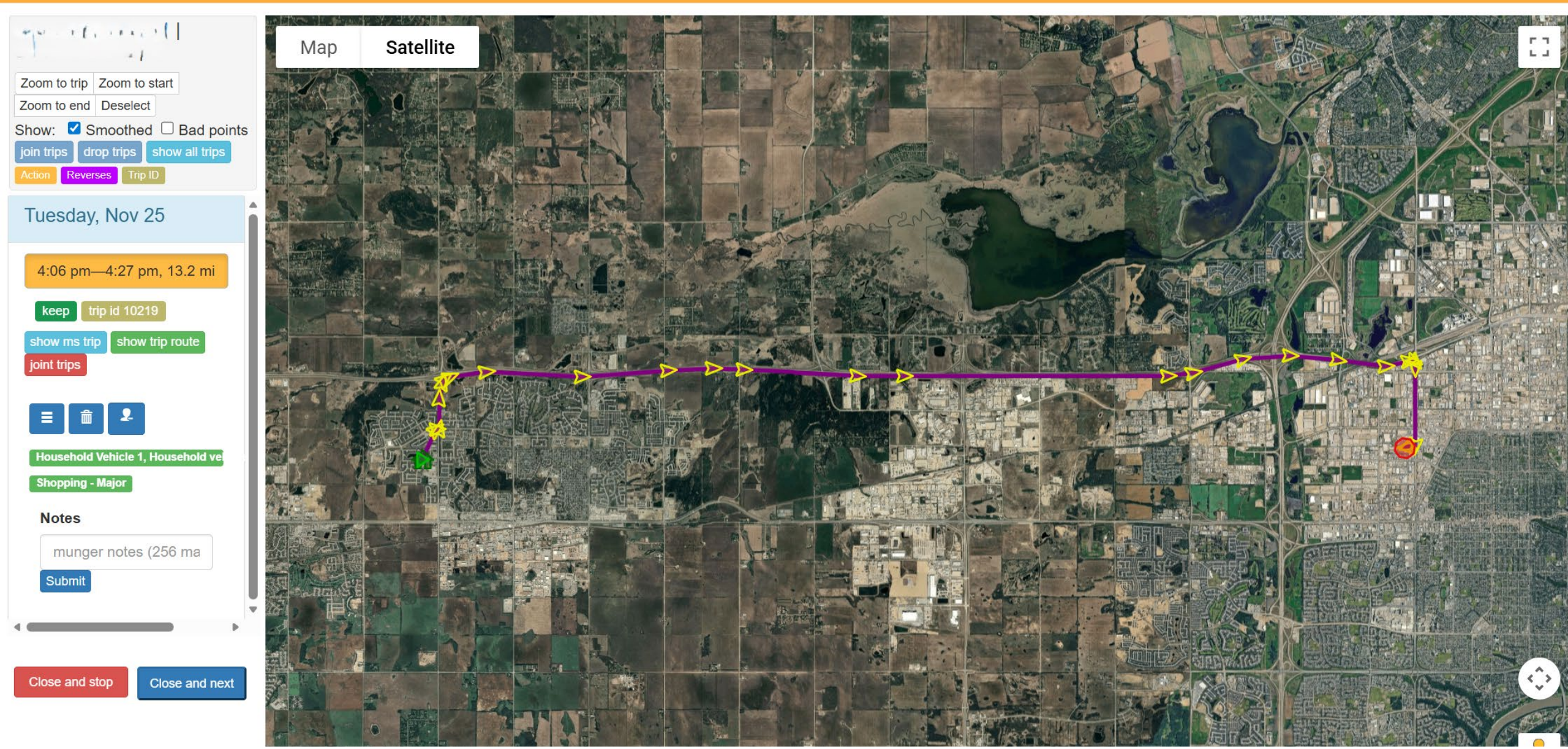
Manual Data Validation

Analysts review flagged trips to split, merge, clean, or remove incorrect records ensuring data accuracy.

Scripted Consistency Checks

Automated scripts perform 150+ checks verifying consistency across household, person, day, vehicle, and location tables.




Zoom to trip
Zoom to start
Zoom to end
Deselect
Show:
Smoothed
Bad points
join trips
drop trips
show all trips
Action
Reverses
Trip ID
Tuesday, Nov 25
4:06 pm—4:27 pm, 13.2 mi
keep
trip id 10219
show ms trip
show trip route
joint trips
Household Vehicle 1, Household ve
Shopping - Major
Notes
munger notes (256 ma
Submit
Close and stop
Close and next
Map
Satellite

Trip Purpose Cleaning

Detect

Identify travel data with trip purpose inconsistencies, using a structured sequence of logical tests (see list), as likely reporting errors.

Remediate

Fix obvious purpose inconsistencies and align destinations with well-known places (home, work, school) using a spatial buffer (~250 m).

Tests

- Check destination purpose and location alignment
- Identify when origin/destinations may have been switched
- Identify potential mis-sequenced trips
- Resolve missing trips
- Impute missing purposes from other trips in the same location

Final Checks

QA/QC Review Processes

Final QA/QC reviews confirm consistent variable naming and prevent data loss during table joins.

Variable Summaries and Validation

Summaries of key variables validate internal consistency and support independent reviews before modeling.

Trip Data Consistency Checks

Checks include totals and distributions of trips across households, persons, days, modes, and residence factors.

Ready Dataset for Modeling

Final checks ensure datasets are coherent, documented, and ready for regional transportation modeling and planning.

Weighting Methodology

Weighting Objectives and Overview

Objectives

Representativeness

Expanding the survey data to represent the entire region across multiple demographic lines

Mitigate survey bias

Identifying and adjusting from biases arising from non-response and different survey modes

Census Reference Data for Base Weights and Targets

Estimates	5-Year American Community Survey (ACS)	1-Year Public Use Microdata Sample (PUMS)
When Were Data Collected?	Every 5 years	Every year
Weighting Purpose	• Spatially allocate PUMS data in alignment with weighting zones. • Calculate base weights at the sample segment level.	• Calculate adjusted weights at the weighting zone group level. • Calculate weighting target estimates. • Demographic crosstabulations.
Smallest Geographic Unit	Block Group	PUMA

Base Weights: Correcting for Sampling Effort and Response

The base weight is calculated as the inverse of the probability of being included.

$$Base\ Weight_{segment} = 1\ /P_{inclusion} = 1\ /(P_{selection} \times P_{response})$$

$$P_{selection} = \frac{Number\ of\ mailed\ invites}{Total\ number\ of\ households}$$

$$P_{response} = \frac{Number\ of\ responses}{Number\ of\ mailed\ invites}$$

Example:

Sample Segment	Census Households	Survey Completes	Base Weight
Orange: General	953,782	359	2,656.8
Orange: Hard-to-Reach	162,406	61	2,662.4

Base weights correct for unequal chances of inclusion at the sample segment level before demographic weighting begins.

Round 1 Weighting: Adjusting for Demographic Non-Response

The first round of weighting corrects *who* is represented in the survey

What problem does this address?

Some demographic groups are less likely to respond

Base weights alone do not correct for this imbalance

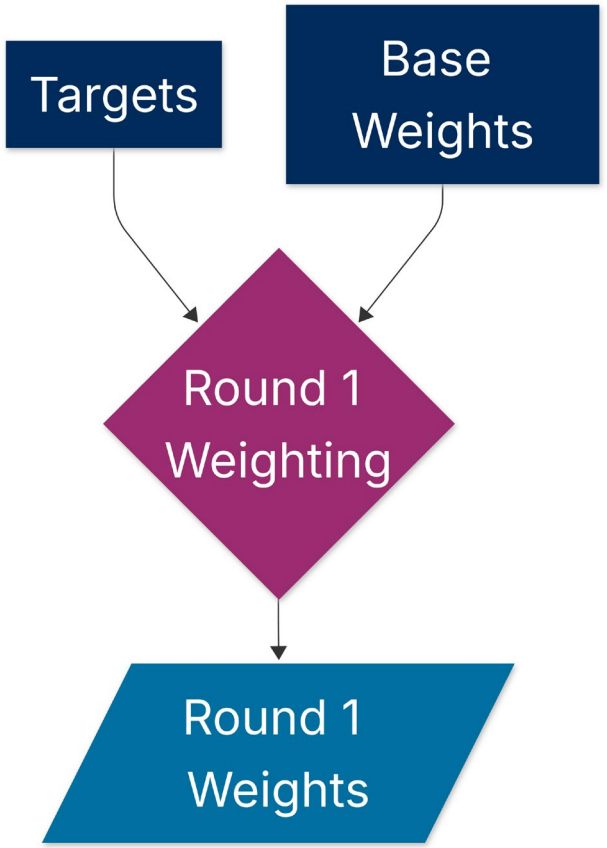
What Round 1 does

Uses **PopulationSim**

Adjusts base weights to match known population demographics

Uses household- and person-level Census targets

Corrects for demographic non-response bias across weighting zones



Round 2 Weighting: Correcting Day-Pattern Reporting Bias

Round 1 corrects who responds; Round 2 corrects how much travel is reported.

Even with demographic correction, not all travel days are equally likely to be reported

Bias depends on survey modality (online, rMove)

Goal: Estimate what *should* be observed if all days were equally likely to be reported

Method: model day-pattern types:

- no-trips
- mandatory trips
- non-mandatory travel only
- children's travel days

Uses Round 1-weighted survey data

Accounts for demographics, household characteristics, survey mode and interactions

Output: Expected daily trip rates by group → **Day-pattern targets** for next round of PopulationSim

Calculating Person, Day, and Trip Weights

From household weights down

Person weights: each person is assigned their household's weight

unrelated household member weights are set to 0 and remaining members weights adjusted up

Day weights: person weight divided by the number of completed days in the household

Unlinked trip weights: each trip is assigned their corresponding day weight

Linked-trip weights: mean of trip weights on the linked-trip

Tour weights: mean of linked-trip weights on the tour

Multi-day adjustment

- Smartphone respondents providing multi-day travel diaries
- Online and call center respondents provide a single-day travel diary
- This adjustment divides day weights by the number of complete, weighting-eligible travel-days

Round 3: Trip-rate Non-response Adjustments

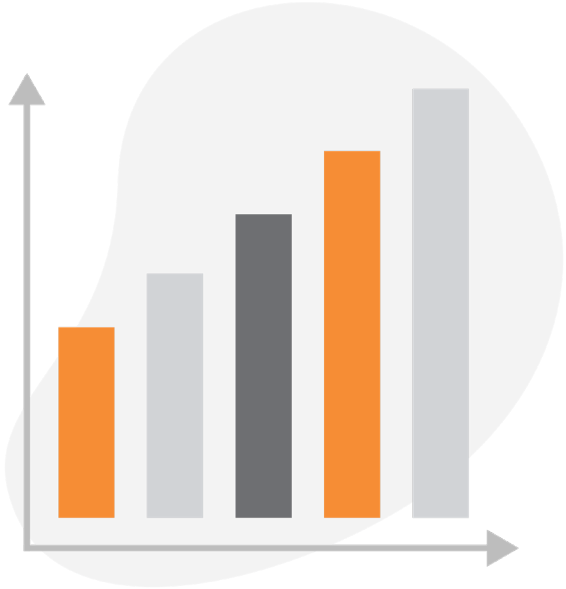
Round 3 accounts for biases in trip rates by diary platform

Problem: Adjusting the weights for day-pattern biases reduces the discrepancy in trip rates between methods but does not eliminate it altogether.

The difference in trip rates tends to be higher for:

- Non-mandatory trips than for mandatory trips, as respondents are less likely to omit their work and school trips in diary-based data.
- Non-home-based trips, since diary respondents often tend to omit intermediate stops on multi-stop tours.

Solution: Model rates across diary platforms and create a factor to boost the lower rates



Weight Validation: An Overview

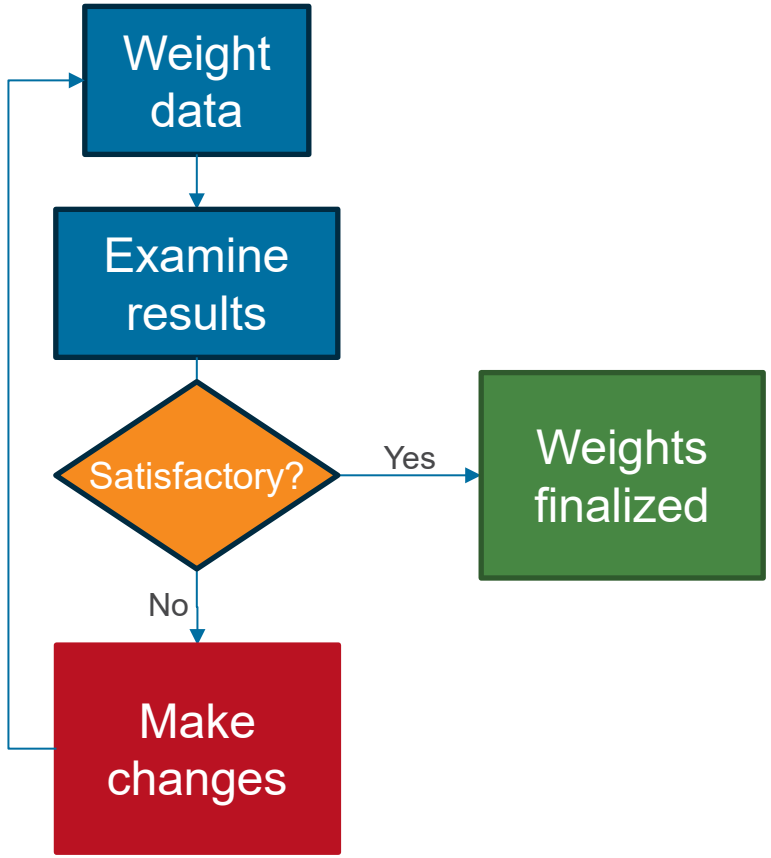
What validation means in weighting

Examine and adjust results to achieve balance between the goals of weighting:

Weighted data represent target population demographic distribution

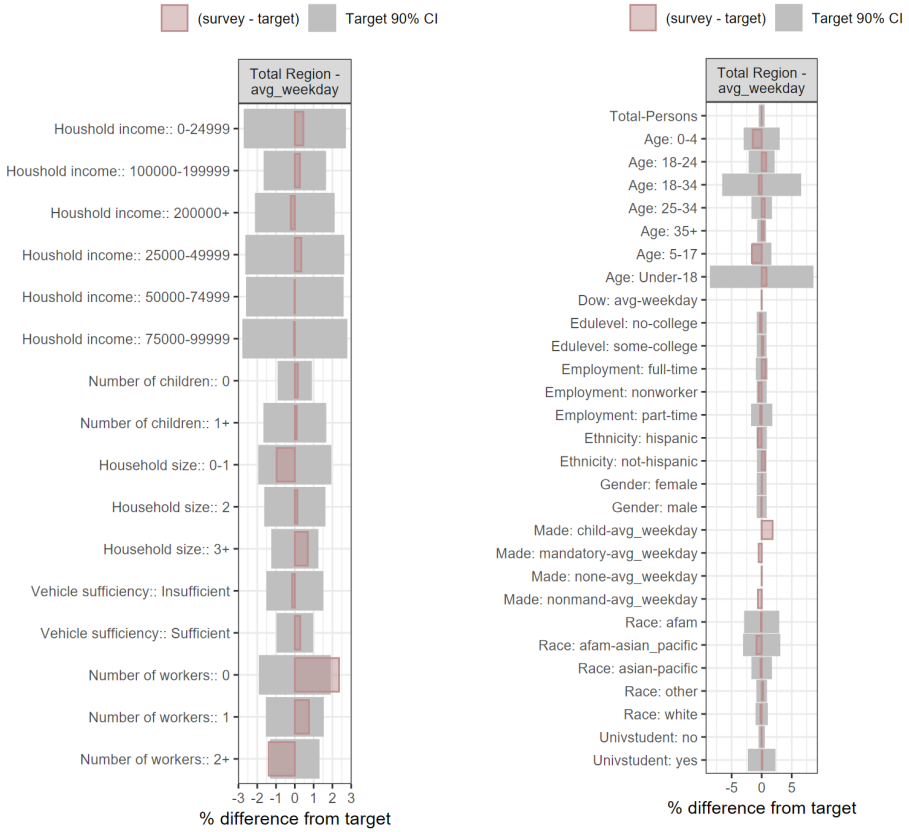
Weighted analyses produce statistically reliable results

What this looks like in practice

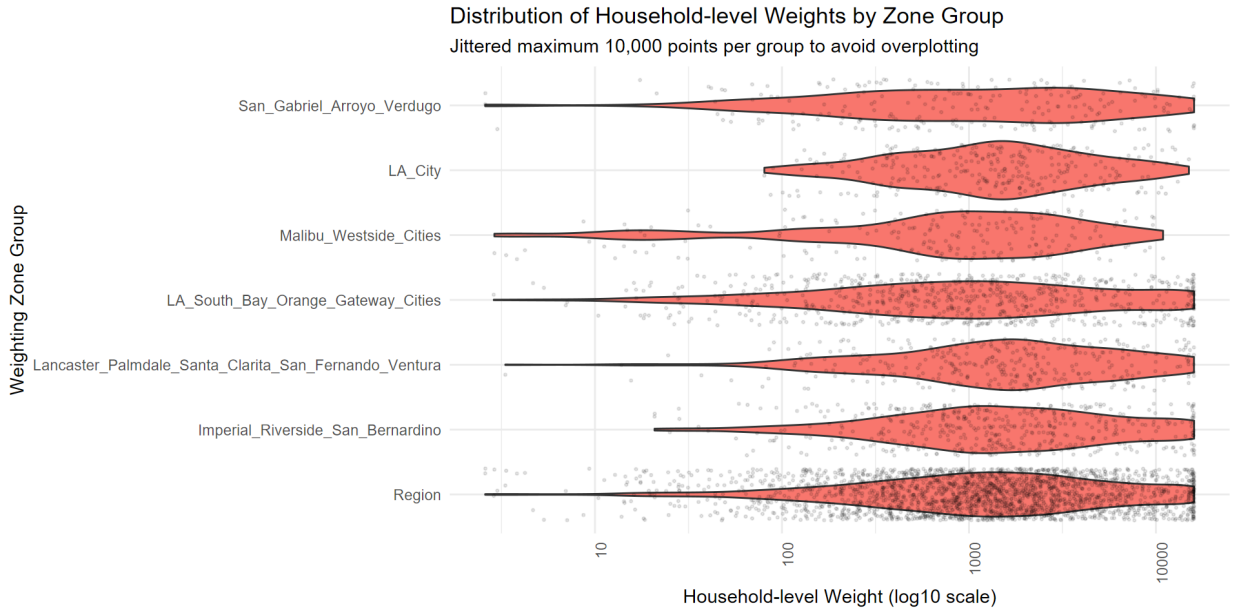


How We Evaluate Weights: Weight Fit and Distribution

Fit-to-targets



Distribution



Target Refinement: Adjusting for Small Sample Sizes

Target level aggregation

- Households with no children are heavily over-represented in this sample → categories 1, 2, and 3+ should be combined, consider dropping the target
- African American and Asian or Pacific Islander respondents are sparse → aggregating these together or into "Other" for more stable estimates

TARGET		Survey Count	Share
Number of children	None	200	86.6%
	1	20	8.7%
	2	10	4.3%
	3 or more	1	0.4%
Race	African American	25	4.5%
	Asian or Pacific Islander	25	4.5%
	Other	250	45.5%
	White	250	45.5%

TARGET		Census Estimate	Share
Number of children	None	500,000	58.8%
	1	200,000	23.5%
	2	100,000	11.8%
	3 or more	50,000	5.9%
Race	African American	200,000	10.8%
	Asian or Pacific Islander	400,000	21.6%
	Other	750,000	40.5%
	White	500,000	27.0%

Presented counts are for illustration purposes only and do not represent the true targets used or counts from the survey data

Target Refinement: Adjusting for Small Sample Sizes

Target level aggregation

- Households with no children are heavily over-represented in this sample → categories 1, 2, and 3+ should be combined, consider dropping the target
- African American and Asian or Pacific Islander respondents are sparse → aggregating these together or into "Other" for more stable estimates

TARGET		Survey Count	Share
Number of children	None	200	86.6%
	1	20	8.7%
	2	10	4.3%
	3 or more	1	0.4%
Race	African American	25	4.5%
	Asian or Pacific Islander	25	4.5%
	Other	250	45.5%
	White	250	45.5%

TARGET		Census Estimate	Share
Number of children	None	500,000	58.8%
	1	200,000	23.5%
	2	100,000	11.8%
	3 or more	50,000	5.9%
Race	African American	200,000	10.8%
	Asian or Pacific Islander	400,000	21.6%
	Other	750,000	40.5%
	White	500,000	27.0%

Presented counts are for illustration purposes only and do not represent the true targets used or counts from the survey data

Target Refinement: Adjusting for Small Sample Sizes

Target level aggregation

- Households with no children are heavily over-represented in this sample → categories 1, 2, and 3+ should be combined, consider dropping the target
- African American and Asian or Pacific Islander respondents are sparse → aggregating these together or into "Other" for more stable estimates

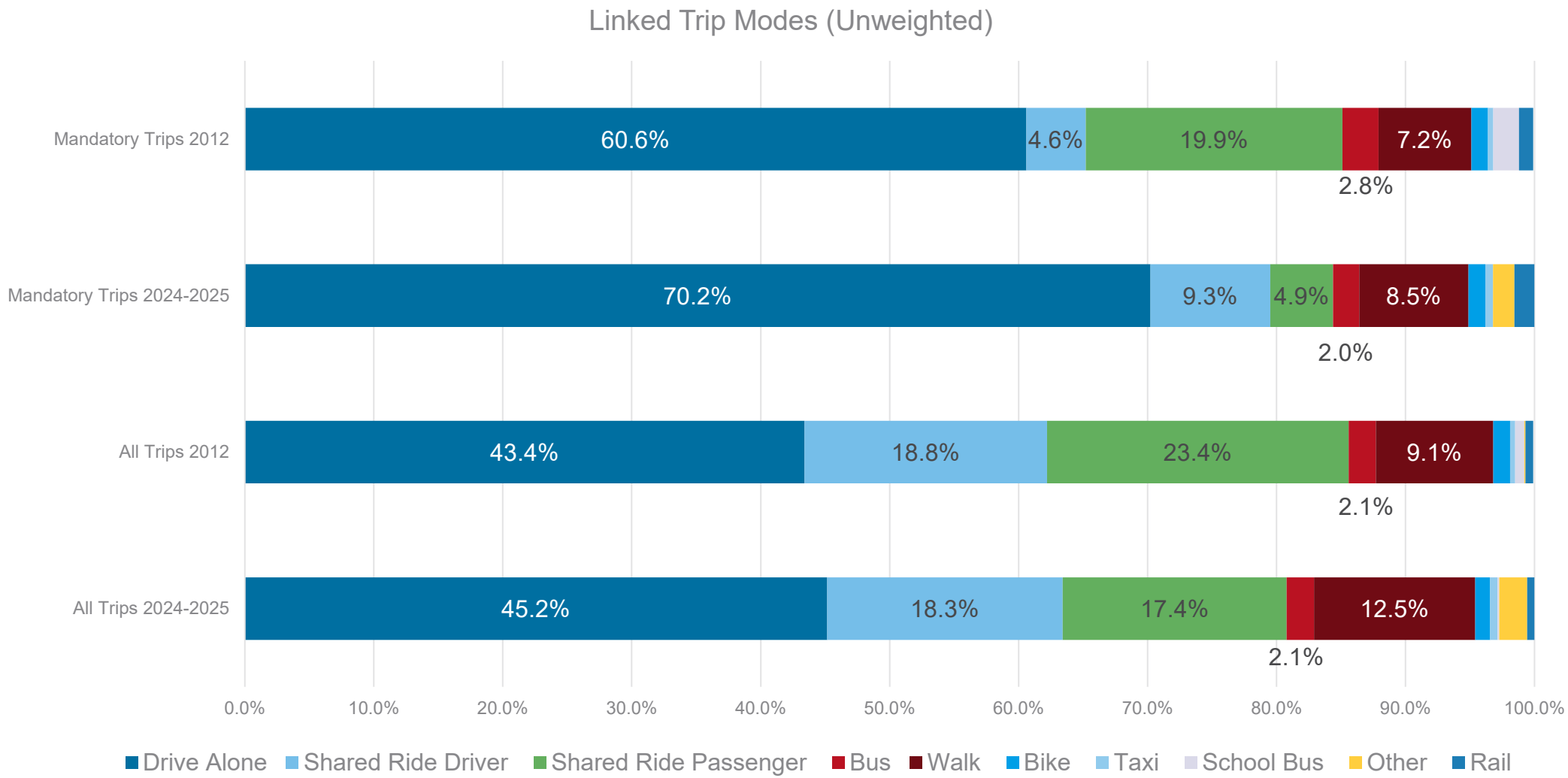
TARGET		Survey Count	Share
Number of children	None	200	86.6%
	1	20	8.7%
	2	10	4.3%
	3 or more	1	0.4%
Race	African American	25	4.5%
	Asian or Pacific Islander	25	4.5%
	Other	250	45.5%
	White	250	45.5%

TARGET		Census Estimate	Share
Number of children	None	500,000	58.8%
	1	200,000	23.5%
	2	100,000	11.8%
	3 or more	50,000	5.9%
Race	African American	200,000	10.8%
	Asian or Pacific Islander	400,000	21.6%
	Other	750,000	40.5%
	White	500,000	27.0%

Presented counts are for illustration purposes only and do not represent the true targets used or counts from the survey data

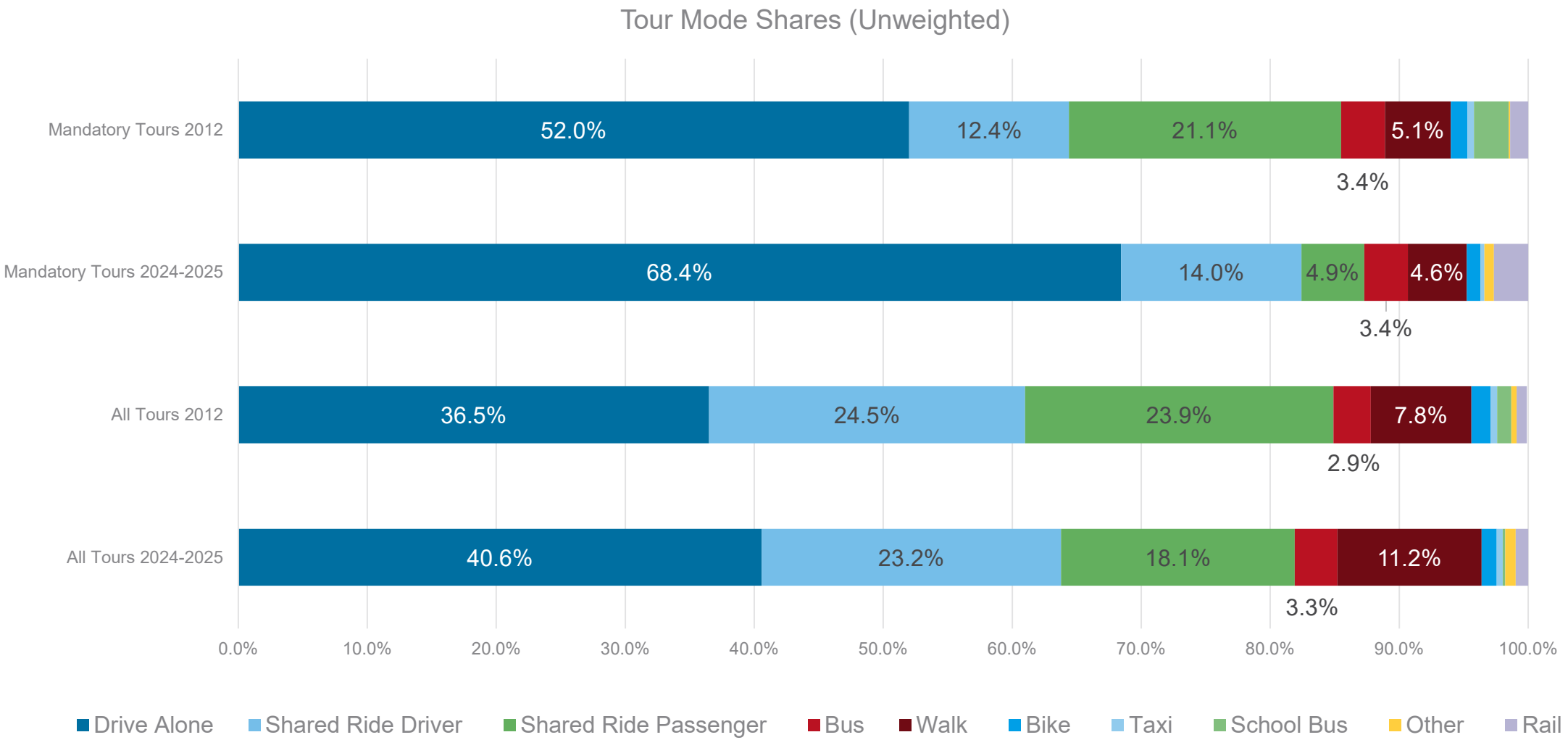

Preliminary
Results

Initial Results

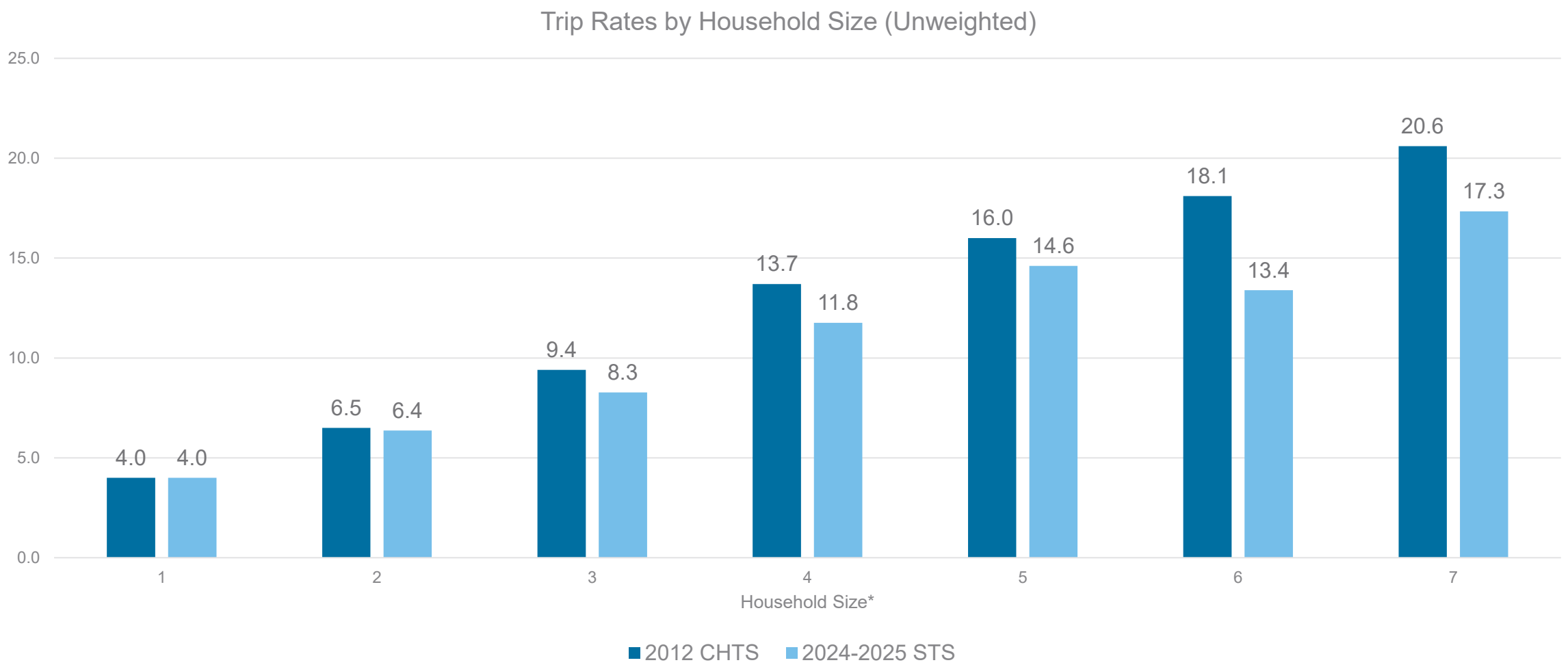


*These are preliminary results based on unweighted data.

Initial Results

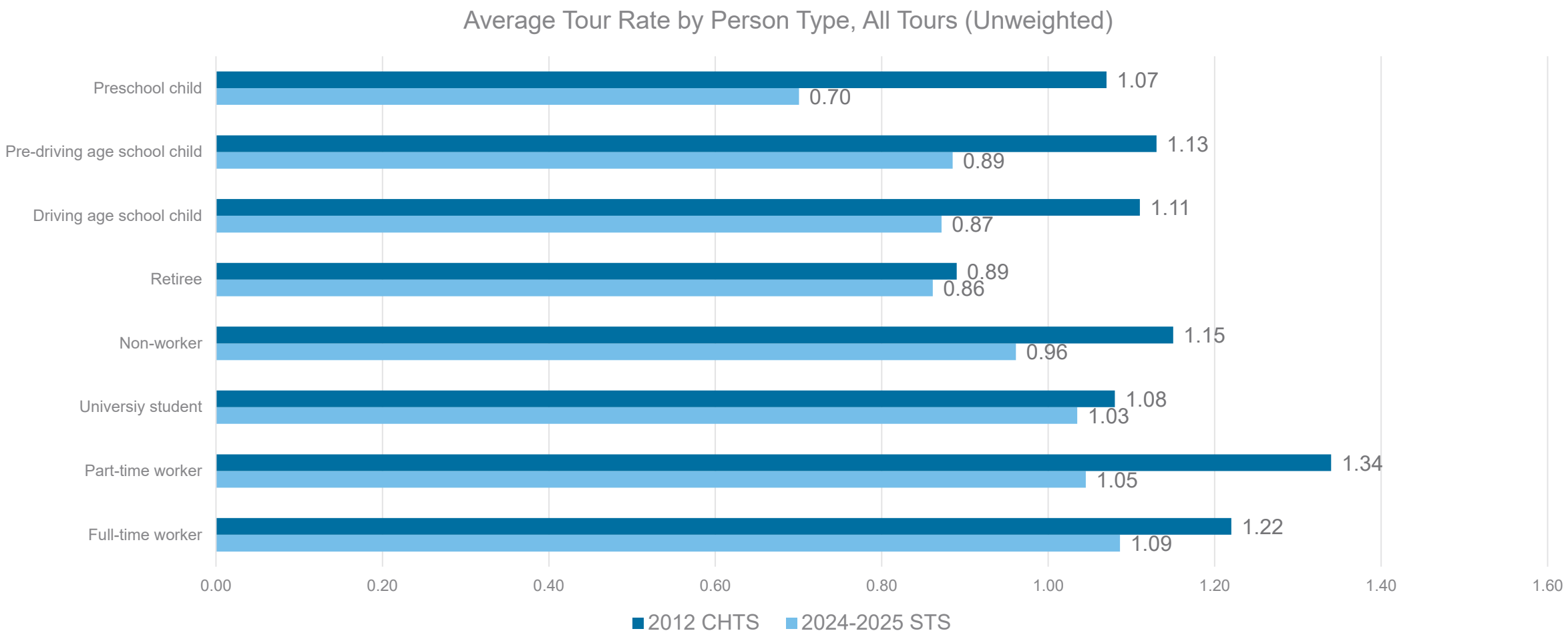


Initial Results

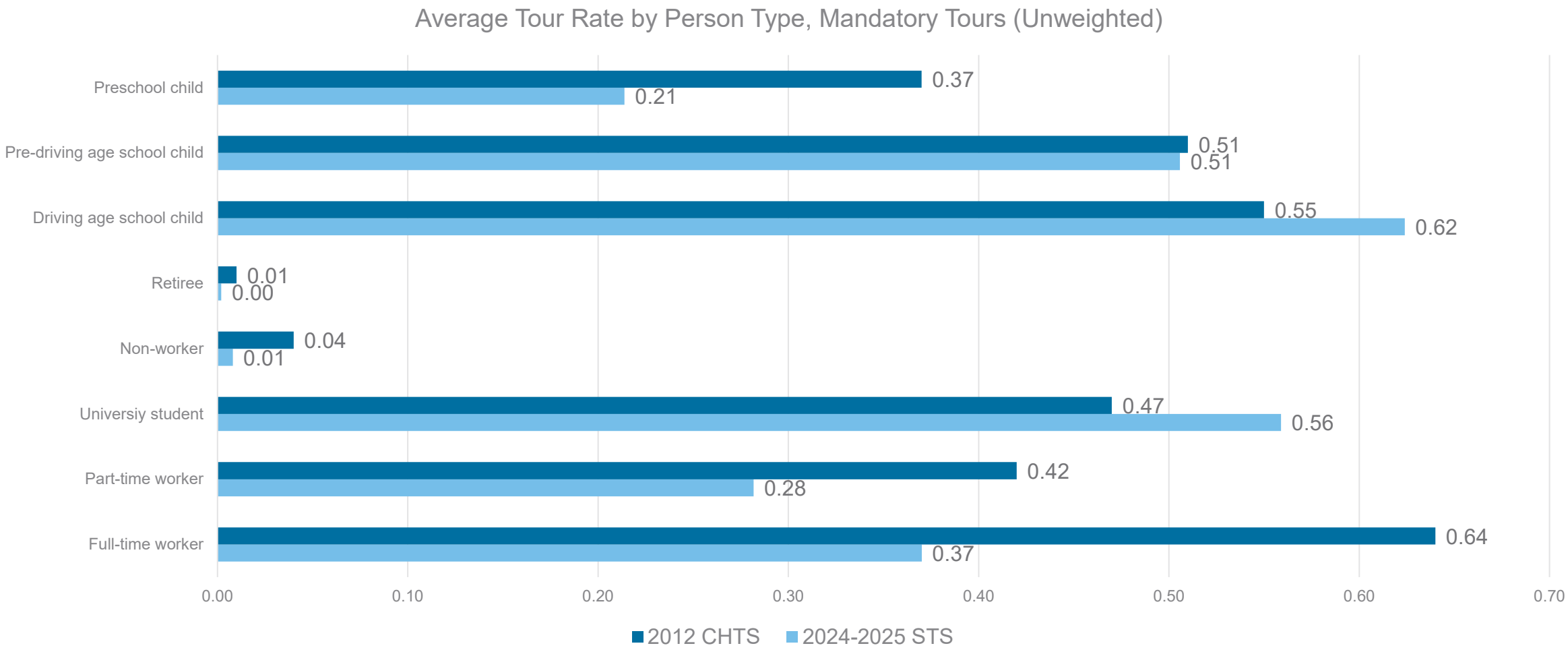


*This chart excludes households larger than 7 people for comparison purposes.

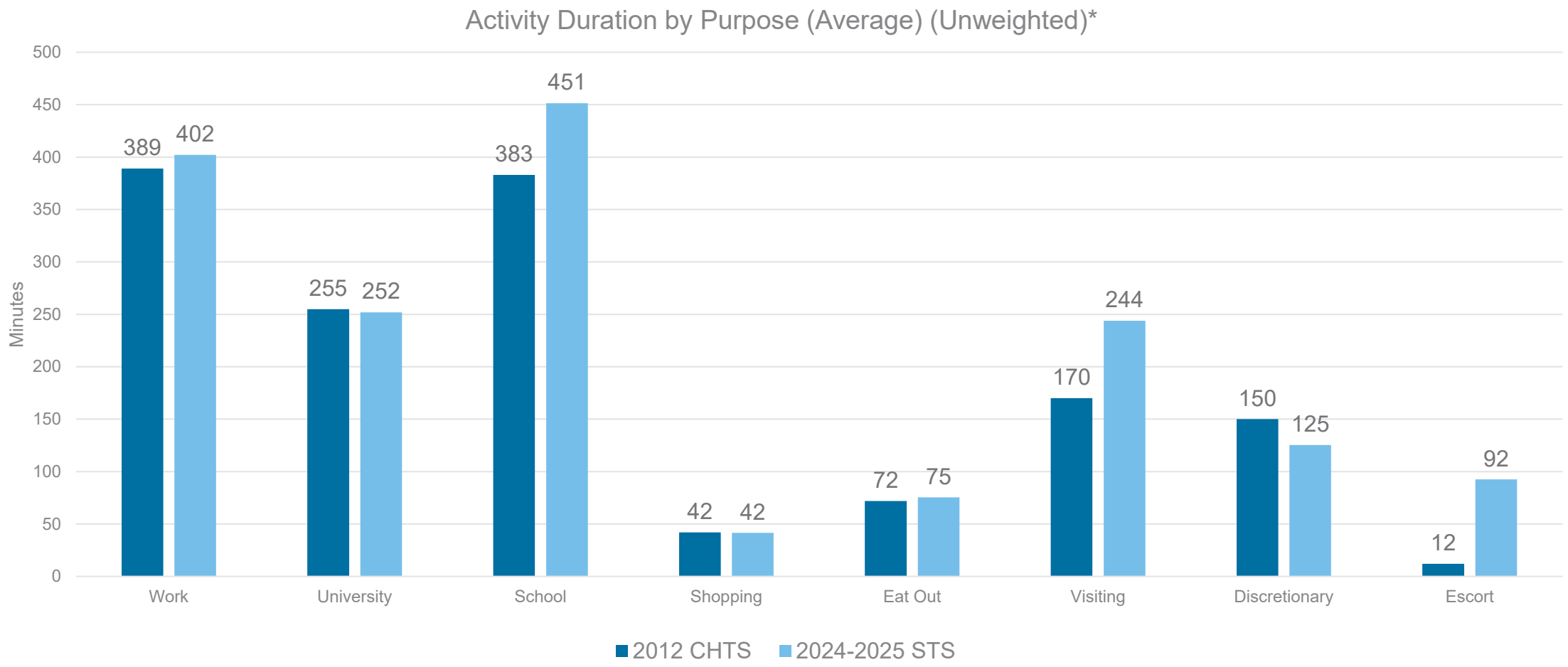
Initial Results



Initial Results



Initial Results



*This chart is not an exhaustive list of all purposes collected in the survey.


Questions?

Contacts

YANG WANG
SENIOR MODELER
wangy@scag.ca.gov

www.rsginc.com