

Narrowbody vs Widebody Fleet Deployment on U.S. Domestic Routes

DATA 407 – Sampling and Design, Winter 2026

Ronit Buti (80223837)

April 2026

1 Introduction

Widebody (twin-aisle) aircraft account for approximately 1.4% of domestic segment records in the United States, yet their deployment is far from random. As route distance increases, so does the presence of widebody equipment, reflecting operational constraints, passenger demand for premium cabins, and carriers' fleet utilization strategies. At the same time, the newest generation of long-range narrowbodies, notably the Airbus A321neo and the forthcoming A321XLR, is expanding the range envelope of single-aisle jets and raising the question of whether widebodies are being displaced on traditionally twin-aisle routes. American Airlines, for example, has 50 A321XLRs on order to replace its premium-configured A321T narrowbodies on transcontinental routes, potentially displacing widebody equipment in the process (Aviation A2Z, 2025), while Delta increased its domestic widebody flying by 36% in summer 2025 even as competitors pulled back (Simple Flying, 2025).

This project applies stratified random sampling to the Bureau of Transportation Statistics (BTS) T-100 Domestic Segment dataset, a complete census of nonstop domestic flights reported monthly by U.S. certificated carriers (BTS, 2025). The full 2025 calendar year file contains $N = 323,364$ segment records spanning 27 carriers, 423 airports, and 13,720 directional routes with nonstop distances ranging from 6 to 5,095 miles. The variable of interest is binary: whether a segment record involves a widebody aircraft ($y_i = 1$) or not ($y_i = 0$), with a population proportion of $p = 0.0139$. The core question is: *at what route distance does the domestic fleet mix shift from predominantly narrowbody to significant widebody presence, and does stratified sampling offer meaningful efficiency gains when estimating a rare-event proportion like this one?*

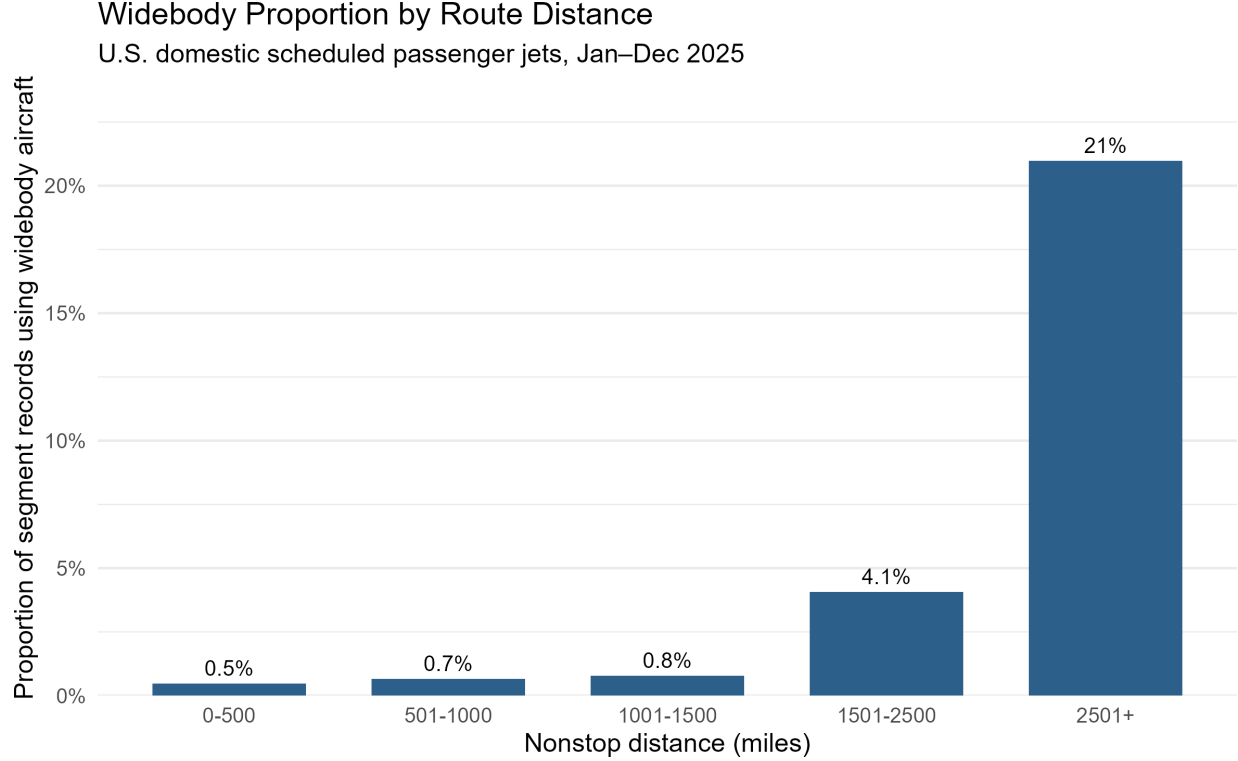


Figure 1: Widebody proportion by route distance band. The jump above 1,500 miles motivates the stratification design.

2 Sampling Design and Methods

Because the widebody proportion varies dramatically across distance bands, from 0.5% on routes under 500 miles to 21.0% on routes exceeding 2,500 miles (Figure 1), a simple random sample would be dominated by the short-haul majority and would rarely capture long-haul widebody operations. Stratified random sampling addresses this by dividing the population into $H = 5$ non-overlapping strata based on nonstop distance (Lohr, 2010). The breakpoints at 500, 1,000, 1,500, and 2,500 miles reflect operational categories in U.S. domestic aviation: short-haul hub shuttles and regional routes (0–500 mi), medium-haul trunk routes (501–1,000 mi), long-haul corridors where narrowbodies dominate but widebodies begin to appear (1,001–1,500 mi), transcontinental routes where fleet mix competition intensifies (1,501–2,500 mi), and ultra-long routes to Hawaii and Alaska where widebodies are most prevalent (2,501+ mi). Within each stratum, a simple random sample is drawn without replacement.

Two allocation strategies are compared for a fixed total sample size of $n = 1,000$ (Shi, 2026b). **Proportional allocation** sets $n_h = n \cdot N_h/N$, sampling each stratum in proportion to its population size. While straightforward, proportional allocation underperforms here because it allocates the bulk of the sample (674 of 1,000 observations) to the two shortest strata, where the widebody

proportion is near zero and within-stratum variance is very low. **Neyman (optimal) allocation** (Neyman, 1934) sets $n_h \propto N_h S_h$, directing more sampling effort toward strata with higher variability, which minimizes $\text{Var}(\hat{p}_{\text{str}})$ for a fixed total n . Since the widebody standard deviation S_h increases sharply with distance ($S_h = 0.069$ for the shortest stratum vs. $S_h = 0.407$ for the longest), Neyman allocation redistributes sample toward the long-haul strata where it is most informative: the 2,501+ mile stratum receives 73 observations under Neyman vs. only 18 under proportional allocation, while the 1,501–2,500 mile stratum receives 242 vs. 120 (Figure 2, Appendix A Table 2).

The stratified estimator of the population proportion is $\hat{p}_{\text{str}} = \sum_{h=1}^H W_h \hat{p}_h$, where $W_h = N_h/N$ is the stratum weight and $\hat{p}_h$ is the sample proportion in stratum h (Shi, 2026a). Its estimated variance, incorporating the finite population correction, is $\widehat{\text{Var}}(\hat{p}_{\text{str}}) = \sum_{h=1}^H (1 - n_h/N_h) W_h^2 s_h^2/n_h$, where s_h^2 is the sample variance within stratum h (as distinct from the population standard deviation S_h used for designing the Neyman allocation above). For comparison, the SRS estimator is $\hat{p} = \bar{y}$ with $\text{SE}(\hat{p}) = \sqrt{(1 - n/N) \cdot s^2/n}$. All confidence intervals use the normal approximation at the 95% level ($z_{0.025} = 1.96$).

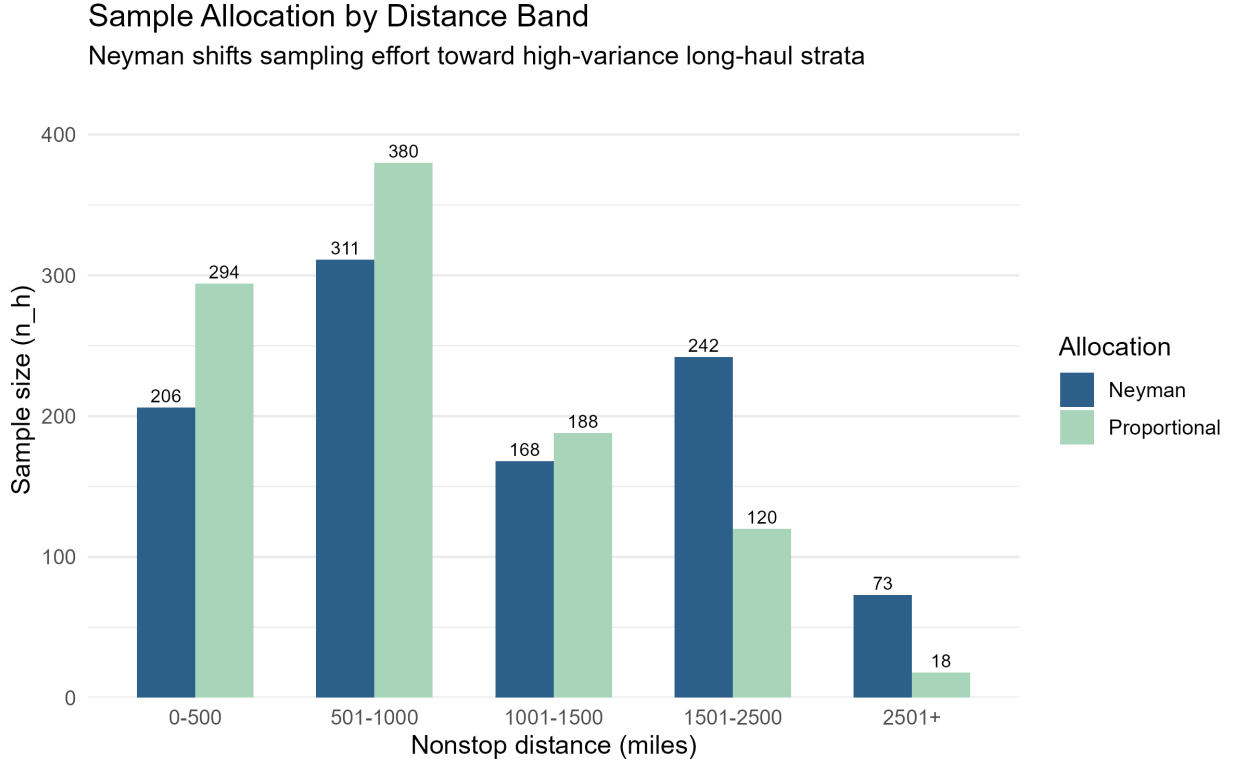


Figure 2: Sample allocation by distance band under proportional vs. Neyman allocation. Neyman shifts effort toward high-variance long-haul strata.

3 Results

Table 1 summarizes the estimation results from a single draw of $n = 1,000$ observations (seed = 654). All three 95% confidence intervals capture the true population proportion $p = 0.0139$. The key finding is the variance reduction achieved by stratification: proportional allocation reduces variance to 91.8% of SRS, while Neyman allocation reduces it to 54.3%, a 46% improvement. This translates to a confidence interval width of 0.0118 for Neyman compared to 0.0160 for SRS.

Table 1: Comparison of estimation methods ($n = 1,000$, true $p = 0.0139$).

Method	$\hat{p}$	SE	95% CI	CI Width	Var (% of SRS)
SRS	0.0170	0.0041	[0.0090, 0.0250]	0.0160	100%
Stratified (proportional)	0.0160	0.0039	[0.0084, 0.0237]	0.0153	91.8%
Stratified (Neyman)	0.0128	0.0030	[0.0069, 0.0187]	0.0118	54.3%

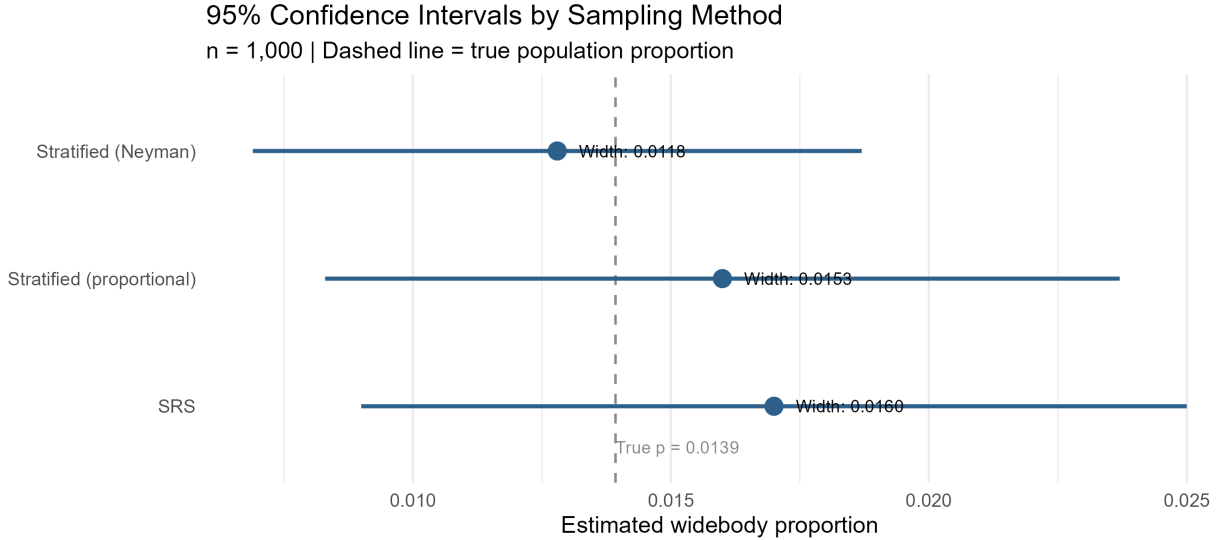


Figure 3: 95% confidence intervals for the three estimation methods. Neyman allocation produces the tightest interval while still capturing the true proportion (dashed line).

At the stratum level, the Neyman estimates reveal the fleet composition gradient in detail (Figure 4). For routes below 1,000 miles, the estimated widebody proportion is under 1%, with tight confidence intervals confirming near-total narrowbody dominance in these strata. The 1,001–1,500 mile band shows a similarly low estimate ($\hat{p}_h = 0.6\%$), though the CI is wider due to the smaller sample allocation. The transition becomes apparent in the 1,501–2,500 mile stratum, where the estimate rises to 2.5% with a CI of [0.5%, 4.4%]. The 2,501+ mile stratum shows the highest estimate at 26.0% with the widest CI ([16.0%, 36.1%]), reflecting the genuine fleet heterogeneity on ultra-long routes where carriers make fundamentally different equipment choices. Importantly, all five stratum-level confidence intervals contain their corresponding true population values, validating

the estimator's performance.

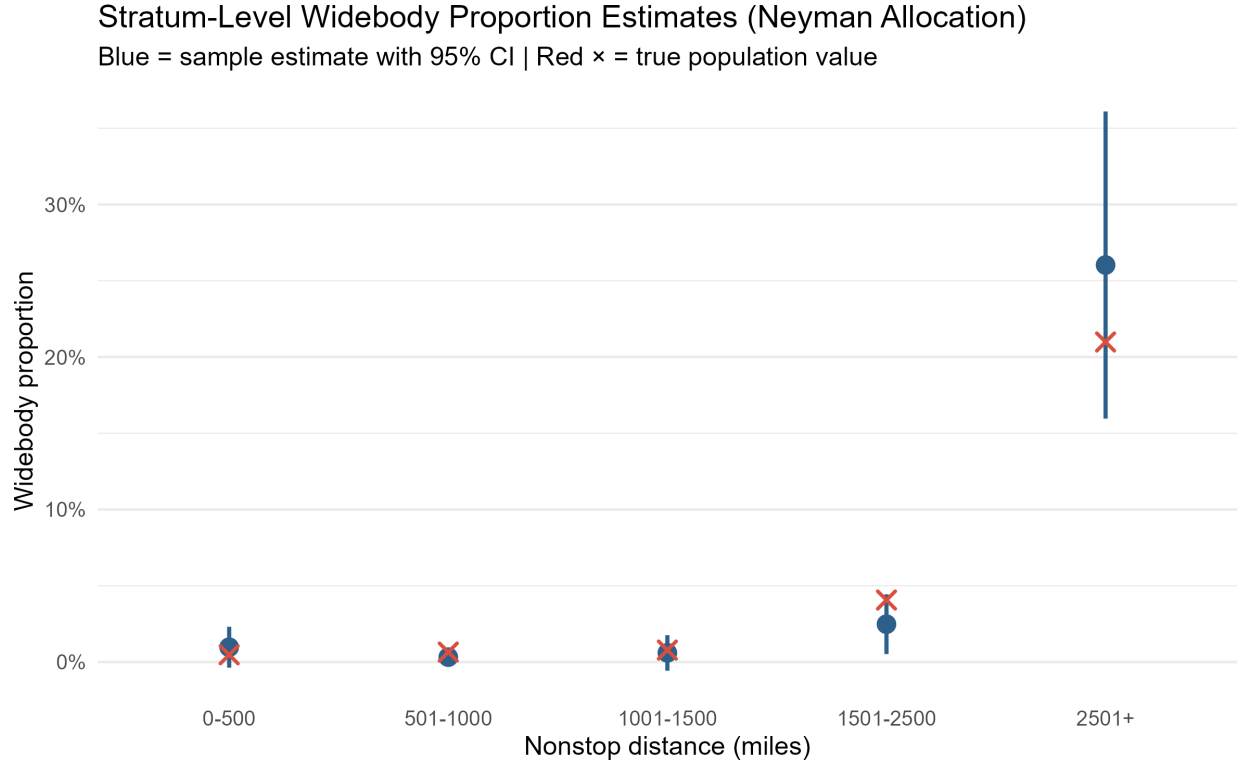


Figure 4: Stratum-level widebody proportion estimates under Neyman allocation. Blue points with bars are sample estimates and 95% CIs; red crosses mark true population values.

To verify that the single-sample results are not artifacts of a particular random draw, a simulation with $B = 2,000$ independent replications was conducted. In each replication, fresh stratified (Neyman and proportional) and SRS samples of $n = 1,000$ were drawn and the population proportion re-estimated. Figure 5 overlays the resulting sampling distributions. All three estimators are unbiased: the mean estimates across replications are 0.01387 (SRS), 0.01389 (proportional), and 0.01394 (Neyman), each within 0.0001 of the true $p = 0.0139$. The Neyman estimator has a standard deviation of 0.00308 across replications, which is 81.5% of the SRS standard deviation (0.00378), confirming that Neyman allocation delivers consistent efficiency gains. The single-sample variance ratio (54.3%) is somewhat more optimistic than the simulation-based ratio of approximately 66%, which is expected: a single draw can over- or underestimate the variance reduction, and the simulation provides the more stable long-run picture.

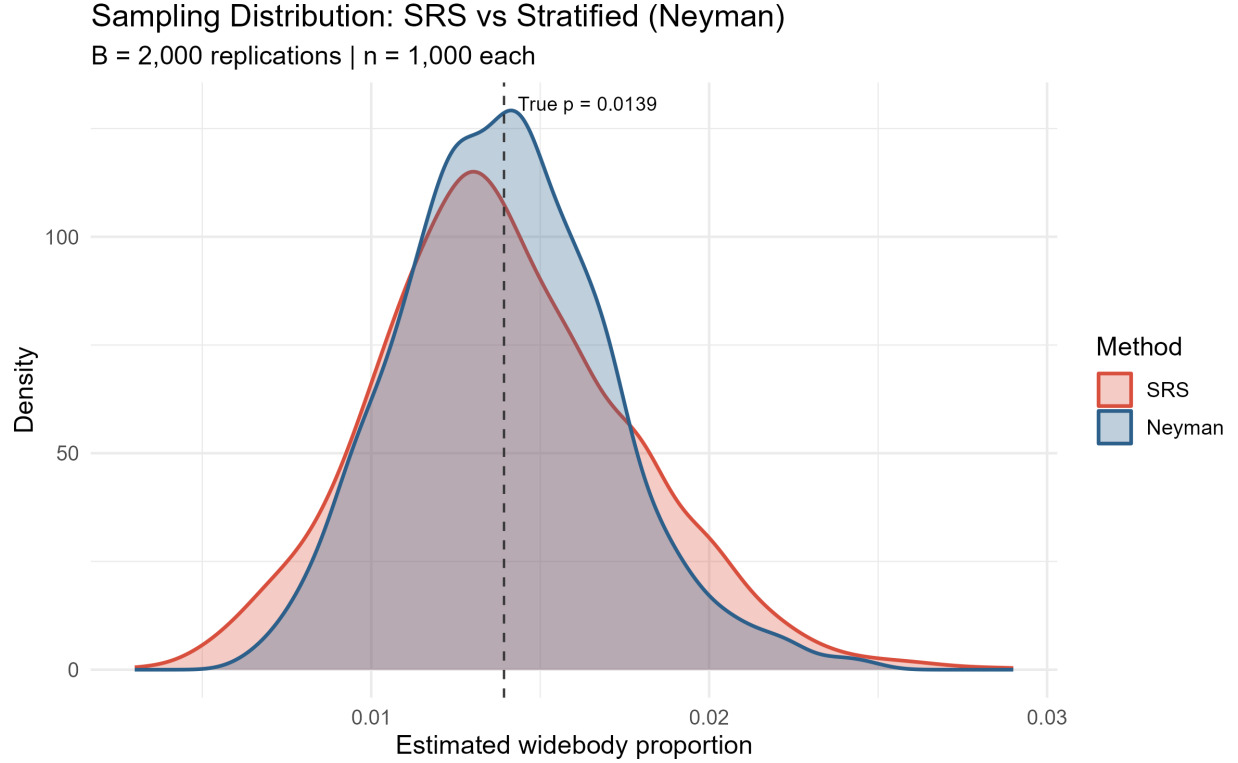


Figure 5: Sampling distributions from 2,000 replications. The Neyman estimator (blue) has a tighter distribution around the true value than SRS (red), confirming the variance reduction.

4 Discussion and Conclusions

This analysis demonstrates that stratified random sampling with Neyman allocation is substantially more efficient than SRS for estimating rare-event proportions in populations with strong structural heterogeneity. With the same total sample size of $n = 1,000$, Neyman allocation reduced estimation variance by 46% relative to SRS, producing a tighter confidence interval that still captured the true widebody proportion. The 2,000-replication simulation confirmed that this gain is not due to sampling luck: the Neyman estimator is consistently more precise across repeated draws.

From an aviation perspective, the data reveal a clear distance threshold around 1,500 miles. Below this distance, widebodies account for less than 1% of segment records and appear primarily as hub repositioning flights, for example, Delta rotating A330s and A350s through ATL–MCO at just 404 miles, where the widebody share reaches 23%. Above 2,500 miles, widebodies comprise roughly one in five segment records. The JFK–LAX corridor (2,475 miles), one of the most prominent domestic widebody routes in the country (Simple Flying, 2024), exemplifies this transition: approximately 47% of segment records involve widebody equipment. Delta deploys 767-400ERs and A330s on the route, while JetBlue and American compete with A321neos and A321XLRs, which actually lead by departure count despite the strong widebody presence. This coexistence of narrowbody and

widebody equipment on the same route illustrates exactly the fleet mix competition that motivates this study.

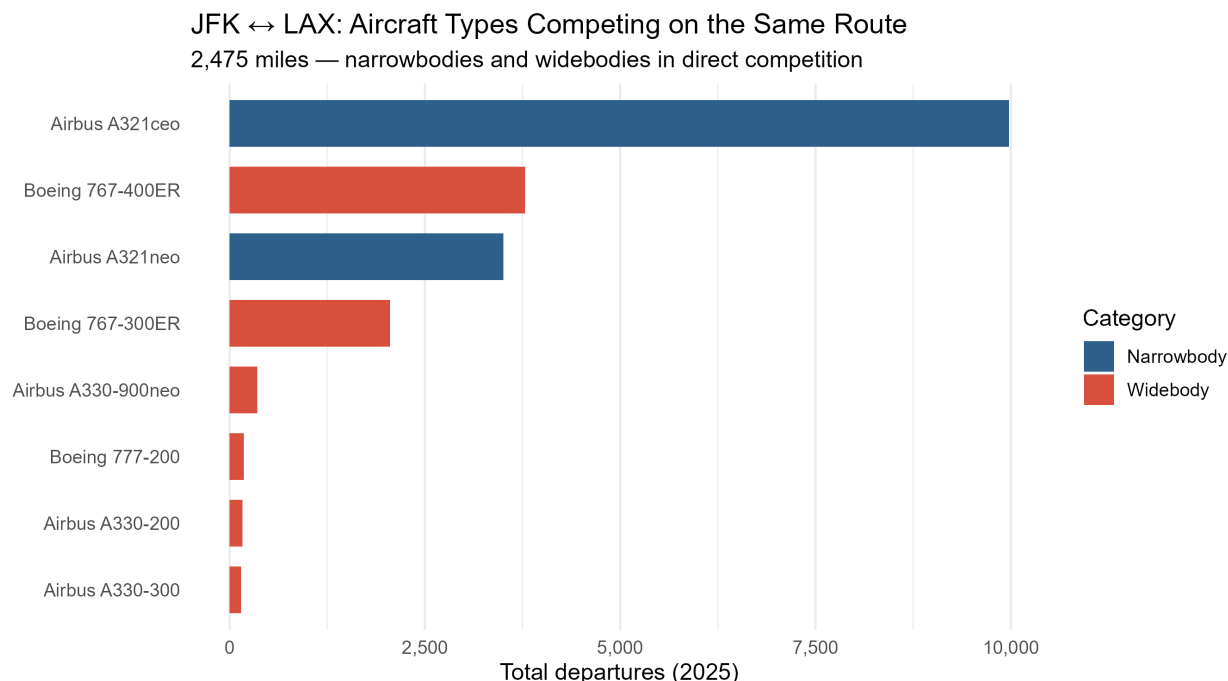


Figure 6: Aircraft types competing on JFK–LAX (2,475 mi). Narrowbodies and widebodies operate side by side on the same transcontinental route.

Carrier strategies diverge significantly on long-haul routes. Delta operates widebodies on 25.2% of its domestic segments over 2,000 miles and notably increased domestic widebody capacity by 36% in summer 2025 (Simple Flying, 2025). Hawaiian Airlines deploys widebodies on 45.7% of long-haul segments, driven by over-water routes to the mainland. In contrast, Alaska Airlines, JetBlue, Southwest, and Frontier operate all-narrowbody fleets on long-haul domestic routes, reflecting both fleet composition and business model differences (Appendix A, Table 3).

These findings are relevant to several audiences. Airline network planners can use distance-stratified fleet mix data to identify corridors where upgrading or downgauging aircraft has the greatest impact on capacity and unit costs. Airport operations teams benefit from knowing where widebody gate and taxiway requirements are concentrated, particularly at hubs like ATL and JFK where widebody domestic flying is disproportionately common. DOT and FAA analysts can track whether the new generation of long-range narrowbodies is shifting competitive dynamics on transcontinental corridors. From a statistical perspective, this project demonstrates that domain knowledge about population structure, here the strong relationship between route distance and fleet mix, can be leveraged through Neyman allocation to achieve substantial efficiency gains over SRS, even for proportions as rare as 1.4%.

5 References

- Aviation A2Z. (2025, February 26). American Airlines Fleet and Network in 2025. *Aviation A2Z*. Retrieved from <https://aviationa2z.com/index.php/2025/02/26/american-airlines-fleet-and-network-in-2025/>
- Bureau of Transportation Statistics. (2020). *Reference File: DB28 Segment Data Product*. U.S. Department of Transportation. Retrieved from <https://www.bts.gov/sites/bts.dot.gov/files/docs/explore-topics-and-geography/topics/airlines-and-airports/230176/reference-file-db28-segment-data-product.pdf>
- Bureau of Transportation Statistics. (2025). *Data Bank 28DS: T-100 Domestic Segment Data (World Area Code)*. U.S. Department of Transportation. Retrieved from <https://www.bts.gov/browse-statistical-products-and-data/bts-publications/data-bank-28ds-t-100-domestic-segment-data>
- Lohr, S. L. (2010). *Sampling: Design and Analysis* (2nd ed.). Brooks/Cole, Cengage Learning.
- Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4), 558–625.
- Shi, X. (2026a). Lecture 6: Stratified Sampling – Introduction and Theory. DATA 407, University of British Columbia Okanagan.
- Shi, X. (2026b). Lecture 7: Stratified Sampling – Allocation and Optimal Design. DATA 407, University of British Columbia Okanagan.
- Simple Flying. (2024, November 27). The US Domestic Routes Operated By Widebody Aircraft. *Simple Flying*. Retrieved from <https://simpleflying.com/us-domestic-routes-operated-widebody-aircraft/>
- Simple Flying. (2025, April 1). Delta Air Lines Plans 36% More Domestic US Widebody Flights This Summer. *Simple Flying*. Retrieved from <https://simpleflying.com/delta-air-lines-plans-36-more-domestic-us-widebody-flights-this-summer/>

6 Appendix A: Supplementary Figures and Tables

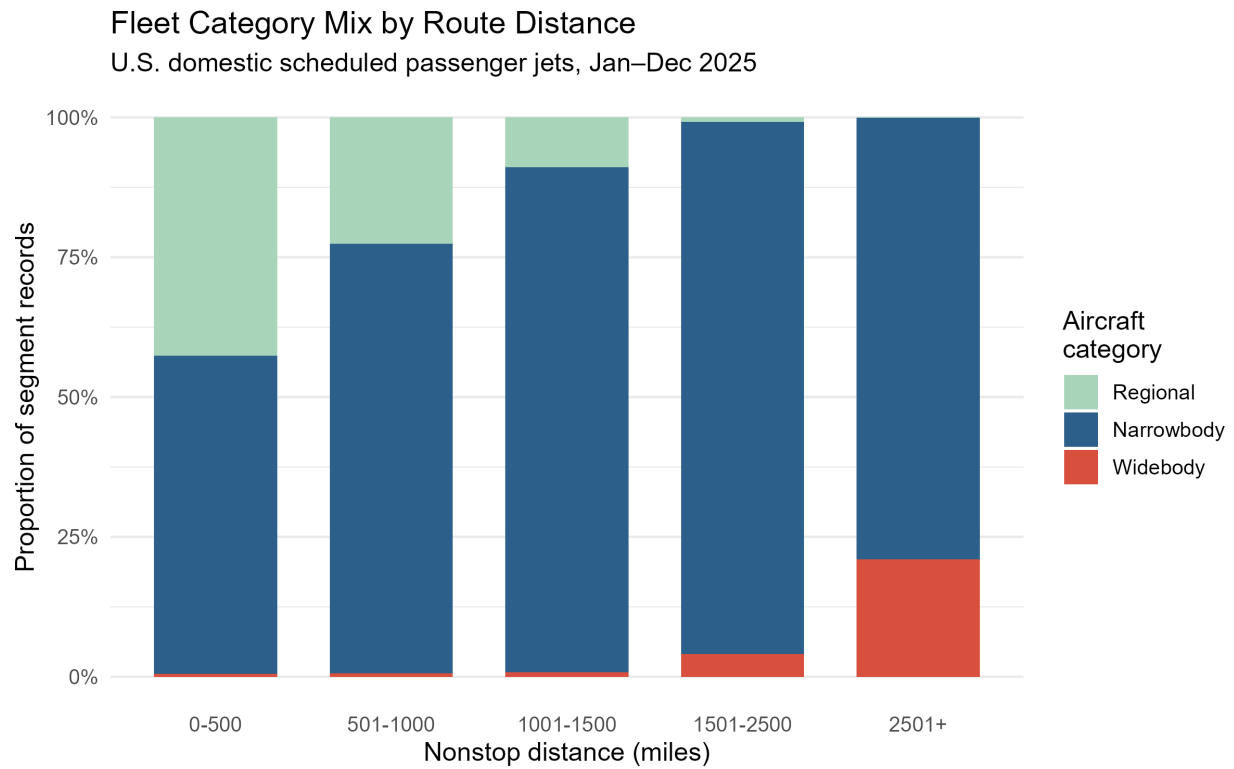


Figure 7: Fleet category mix by distance band. Regional jets dominate short-haul; narrowbodies dominate mid-range; widebodies emerge on the longest routes.

Population Size by Distance Band (Stratum)

Short-haul segments vastly outnumber long-haul — motivates stratification

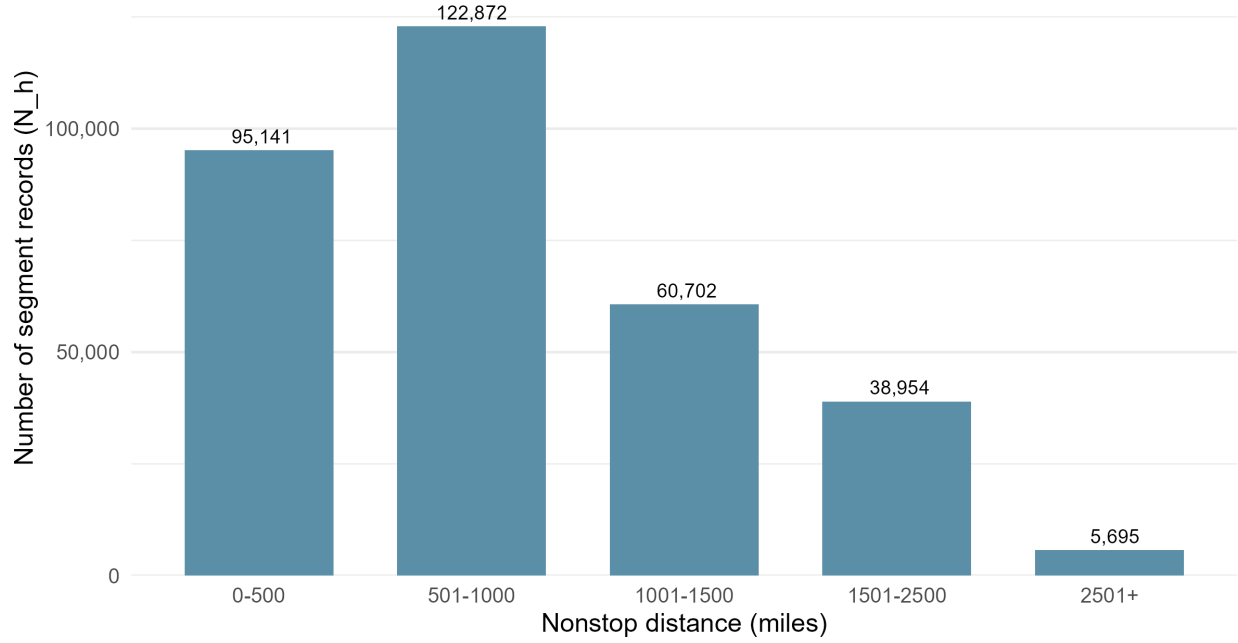


Figure 8: Population sizes by distance band. The heavy concentration of records in short-haul strata motivates stratification to ensure adequate representation of long-haul routes.

Table 2: Population parameters and sample allocation by stratum.

Distance Band	N_h	W_h	p_h	S_h	n (Prop.)	n (Neyman)
0-500	95,141	0.2942	0.0048	0.0688	294	206
501-1000	122,872	0.3800	0.0065	0.0804	380	311
1001-1500	60,702	0.1877	0.0078	0.0881	188	168
1501-2500	38,954	0.1205	0.0406	0.1975	120	242
2501+	5,695	0.0176	0.2098	0.4072	18	73

Table 3: Carrier fleet mix on long-haul domestic routes ($\geq 2,000$ miles).

Carrier	Departures	Widebody %
UA	100,814	17.6
AA	85,828	9.9

Carrier	Departures	Widebody %
DL	82,680	25.2
AS	66,156	0.0
B6	38,615	0.0
WN	29,214	0.0
HA	26,095	45.7
NK	7,759	0.0
F9	4,483	0.0
MX	4,040	0.0

7 Appendix B: R Code

The full analysis pipeline consists of six scripts executed sequentially via `run_all.R`. Source code is reproduced below. The raw data file (`db28seg.dd.wac.202501.202512.asc`) must be downloaded from the BTS Transtats portal and placed in `data/raw/` to reproduce the analysis. All outputs are reproducible with `set.seed(654)`.

7.1 Setup and Constants

```
# =====
# 00_setup.R
# DATA 407 Project - Narrowbody vs Widebody Fleet Deployment
# Ronit Buti (80223837)
#
# Purpose: Load packages, set seed, define shared constants and mappings
# =====

# --- Packages -----
library(data.table)
library(ggplot2)
library(scales)

# --- Reproducibility -----
set.seed(654)

# --- File paths -----
RAW_FILE    <- "data/raw/db28seg.dd.wac.202501.202512.asc"
CLEAN_FILE  <- "data/processed/flights_clean.csv"
FIG_DIR     <- "output/figures"
TBL_DIR     <- "output/tables"

# Create output directories if they don't exist
dir.create(FIG_DIR, recursive = TRUE, showWarnings = FALSE)
dir.create(TBL_DIR, recursive = TRUE, showWarnings = FALSE)
dir.create("data/processed", recursive = TRUE, showWarnings = FALSE)

# --- Column names for raw BTS T-100 Segment file -----
# Based on the DB28 Segment Data Product Reference (BTS, 2020)
# Source: https://www.bts.gov/sites/bts.dot.gov/files/docs/explore-topics-and-
#         geography/topics/airlines-and-airports/230176/reference-file-db28-
#         segment-data-product.pdf
COL_NAMES <- c(
  "Year", "Month",
  "OriginAirport", "OriginNumCode", "OriginWAC", "OriginCity",
```

```

"DestAirport", "DestNumCode", "DestWAC", "DestCity",
"Carrier", "CarrierEntity", "CarrierGroup",
"Distance",      # great-circle statute miles (BTS-calculated)
"Class",         # F=sched pax/cargo, G=sched all-cargo, L=charter pax, P=charter
  ↪ cargo
"AircraftGroup", # 6=2-eng jet, 7=3-eng jet, 8=4-eng jet, 4/5=turboprop, etc.
"AircraftType",  # 3-digit code for specific make/model
"AircraftConfig", # 1=pax, 2=cargo, 3=combi, 4=amphibious
"DepPerformed", "DepScheduled",
"PayloadLbs", "AvailSeats",
"Passengers", "Freight", "Mail",
"RampMinutes", "AirborneMinutes",
"CarrierWAC"    # carrier's country (10 = U.S.)
)

# --- Aircraft type classification -----
# Maps BTS AircraftType codes to specific models and body categories.
#
# Verification method: Each code was cross-referenced against (1) the carriers
# reporting it, (2) median seats per departure, and (3) known fleet compositions
# from airline fleet pages.
#
# Categories:
#   Narrowbody - single-aisle jets (Boeing 737 family, A320 family, 757, 717, A220)
#   Widebody   - twin-aisle jets (767, 777, 787, A330, A350)
#   Regional   - small jets typically <100 seats (CRJ, ERJ, E170/175)
AIRCRAFT_MAP <- data.table(
  AircraftType = c(
    # --- Narrowbody (16 types) ---
    608, 612, 614, 622, 623, 634,      # B717, B737-700, -800, B757-200, -300,
    ↪ B737-900
    694, 698, 699,                    # A320, A319, A321ceo
    721, 722, 723, 724,                # A321neo, A320neo, A220-100, A220-300
    838, 839, 888,                    # B737 MAX 8, MAX 9, B737-900ER
    # --- Widebody (11 types) ---
    624, 626,                         # B767-300ER, -400ER
    627, 637,                         # B777-200, -300ER
    887, 889, 837,                    # B787-8, -9, -10
    696, 687, 339,                    # A330-200, -300, A330-900neo
    359,                              # A350-900
    # --- Regional jet (8 types) ---
    530, 629, 631, 638,                # ERJ-135, ERJ-145, CRJ-200, CRJ-700
    673, 675, 677, 678                # CRJ-900/E175, ERJ-145, E170, E190
  ),

```

```

AircraftName = c(
  # Narrowbody
  "Boeing 717-200", "Boeing 737-700", "Boeing 737-800",
  "Boeing 757-200", "Boeing 757-300", "Boeing 737-900",
  "Airbus A320", "Airbus A319", "Airbus A321ceo",
  "Airbus A321neo", "Airbus A320neo", "Airbus A220-100", "Airbus A220-300",
  "Boeing 737 MAX 8", "Boeing 737 MAX 9", "Boeing 737-900ER",
  # Widebody
  "Boeing 767-300ER", "Boeing 767-400ER",
  "Boeing 777-200", "Boeing 777-300ER",
  "Boeing 787-8", "Boeing 787-9", "Boeing 787-10",
  "Airbus A330-200", "Airbus A330-300", "Airbus A330-900neo",
  "Airbus A350-900",
  # Regional jet
  "Embraer ERJ-135", "Embraer ERJ-145", "CRJ-200", "CRJ-700",
  "CRJ-900 / E175", "Embraer ERJ-145", "Embraer E170", "Embraer E190"
),
Category = c(
  rep("Narrowbody", 16),
  rep("Widebody", 11),
  rep("Regional", 8)
)
)

# --- Distance band definitions -----
# Operationally meaningful strata for U.S. domestic routes
DIST_BREAKS <- c(0, 500, 1000, 1500, 2500, Inf)
DIST_LABELS <- c("0-500", "501-1000", "1001-1500", "1501-2500", "2501+")

# --- Sampling parameters -----
TOTAL_SAMPLE_SIZE <- 1000 # total n across all strata

cat("00_setup.R loaded successfully.\n")
cat("Seed set to 654. Total sample size:", TOTAL_SAMPLE_SIZE, "\n")

```

7.2 Data Cleaning

```

# =====
# 01_clean.R
# DATA 407 Project - Narrowbody vs Widebody Fleet Deployment
# Ronit Buti (80223837)
#
# Purpose: Read raw BTS T-100 segment data, filter to scheduled passenger
#          jet service, classify aircraft as Narrowbody/Widebody/Regional,

```

```

#           assign distance-band strata, and save cleaned dataset.
#
# Input:   data/raw/db28seg.dd.wac.202501.202512.asc
# Output:  data/processed/flights_clean.csv
# =====

source("R/00_setup.R")

# --- 1. Read raw data -----
cat("Reading raw BTS T-100 segment file...\n")
raw <- fread(
  RAW_FILE,
  sep      = "|",
  header   = FALSE,
  col.names = c(COL_NAMES, "Extra"), # trailing pipe creates an empty column
  colClasses = "character"
)
raw[, Extra := NULL] # drop the phantom column

cat("  Raw rows:", format(nrow(raw), big.mark = ","), "\n")

# Convert numeric columns
num_cols <- c("Year", "Month", "Distance", "AircraftGroup", "AircraftType",
              "AircraftConfig", "DepPerformed", "DepScheduled",
              "AvailSeats", "Passengers")
for (col in num_cols) raw[, (col) := as.numeric(get(col))]

# --- 2. Filter to scheduled passenger jet service -----
# Class = "F" : scheduled passenger/cargo (eliminates charter and all-cargo)
# AircraftGroup in {6,7,8} : 2/3/4-engine jets (eliminates turboprops, pistons)
# AircraftConfig = 1 : passenger configuration
cat("Filtering to scheduled passenger jets (Class=F, jet groups, pax config)...\n")
flights <- raw[Class == "F" & AircraftGroup %in% c(6, 7, 8) & AircraftConfig == 1 &
  ↪ Distance > 0]
cat("  Filtered rows:", format(nrow(flights), big.mark = ","), "\n")
cat("  Total departures:", format(sum(flights$DepPerformed, na.rm = TRUE), big.mark =
  ↪ ", "), "\n")

# --- 3. Classify aircraft -----
# Join with our aircraft type mapping
cat("Classifying aircraft types...\n")
flights <- merge(flights, AIRCRAFT_MAP, by = "AircraftType", all.x = TRUE)

# Check coverage

```

```

n_classified <- sum(!is.na(flights$Category))
n_unclassified <- sum(is.na(flights$Category))
cat("   Classified:", format(n_classified, big.mark = ","),
    "(", round(n_classified / nrow(flights) * 100, 1), "%)\n")
cat("   Unclassified:", format(n_unclassified, big.mark = ","),
    "(", round(n_unclassified / nrow(flights) * 100, 1), "%)\n")

# Show unclassified types for transparency
if (n_unclassified > 0) {
  cat("   Unclassified aircraft types (by departures):\n")
  unk <- flights[is.na(Category), .(Deps = sum(DepPerformed)), by = AircraftType]
  print(unk[order(-Deps)][1:10])
}

# Drop unclassified rows (small fraction of data)
flights <- flights[!is.na(Category)]
cat("   After dropping unclassified:", format(nrow(flights), big.mark = ","), "rows\n")

# --- 4. Create analysis variables -----
# Binary widebody indicator (the variable of interest for sampling)
flights[, is_widebody := as.integer(Category == "Widebody")]

# Seats per departure (validation check)
flights[, seats_per_dep := round(AvailSeats / pmax(DepPerformed, 1))]

# Distance band (strata)
flights[, dist_band := cut(Distance,
                          breaks = DIST_BREAKS,
                          labels = DIST_LABELS,
                          right = TRUE)]

# --- 5. Select columns and save -----
keep_cols <- c("Year", "Month", "OriginAirport", "OriginCity",
              "DestAirport", "DestCity", "Carrier", "CarrierGroup",
              "Distance", "dist_band",
              "AircraftType", "AircraftName", "Category", "is_widebody",
              "DepPerformed", "AvailSeats", "Passengers", "seats_per_dep")

flights_out <- flights[, ..keep_cols]

cat("Writing cleaned data to", CLEAN_FILE, "... \n")
fwrite(flights_out, CLEAN_FILE)

# --- 6. Summary -----

```

```

cat("\n==== CLEANING SUMMARY =====\n")
cat("Period: January - December 2025\n")
cat("Rows:", format(nrow(flights_out), big.mark = ","), "\n")
cat("Departures:", format(sum(flights_out$DepPerformed), big.mark = ","), "\n")
cat("Unique carriers:", uniqueN(flights_out$Carrier), "\n")
cat("Unique routes:", uniqueN(paste(flights_out$OriginAirport, flights_out$DestAirport)),
  ↪  "\n\n")

cat("Category breakdown (rows):\n")
print(flights_out[, .(N, Departures = sum(DepPerformed)), by =
  ↪  Category][order(-Departures)])

cat("\nDistance band breakdown (rows):\n")
print(flights_out[, .(N, Departures = sum(DepPerformed)), by =
  ↪  dist_band][order(dist_band)])

cat("\nWidebody proportion by distance band:\n")
wb_by_dist <- flights_out[, .(
  N_h      = .N,
  Widebody = sum(is_widebody),
  Prop     = round(mean(is_widebody), 4)
), by = dist_band][order(dist_band)]
print(wb_by_dist)

cat("\n01_clean.R completed.\n")

```

7.3 Exploratory Data Analysis

```

# =====
# 02_eda.R
# DATA 407 Project - Narrowbody vs Widebody Fleet Deployment
# Ronit Buti (80223837)
#
# Purpose: Population-level exploratory analysis and summary statistics.
#         Since we have a full census (all BTS segment records), we can
#         compute exact population parameters before sampling.
#
# Input:   data/processed/flights_clean.csv
# Output:  output/tables/ - population summary CSVs
#         output/figures/ - EDA plots
# =====

source("R/00_setup.R")

```

```

# --- 1. Load cleaned data -----
cat("Loading cleaned data...\n")
flights <- fread(CLEAN_FILE)
flights[, dist_band := factor(dist_band, levels = DIST_LABELS)]

cat("  Rows:", format(nrow(flights), big.mark = ","), "\n")
cat("  Departures:", format(sum(flights$DepPerformed), big.mark = ","), "\n\n")

# --- 2. Population summary table -----
pop_summary <- data.table(
  Metric = c("Period", "Segment records (N)", "Total departures",
             "Unique carriers", "Unique airports",
             "Unique routes (directional)", "Distance range (mi)"),
  Value = c(
    "January - December 2025",
    format(nrow(flights), big.mark = ","),
    format(sum(flights$DepPerformed), big.mark = ","),
    uniqueN(flights$Carrier),
    uniqueN(c(flights$OriginAirport, flights$DestAirport)),
    uniqueN(paste(flights$OriginAirport, flights$DestAirport)),
    paste0(min(flights$Distance), " - ", max(flights$Distance))
  )
)
cat("=== POPULATION SUMMARY ===\n")
print(pop_summary, row.names = FALSE)
fwrite(pop_summary, file.path(TBL_DIR, "population_summary.csv"))

# --- 3. Category breakdown -----
cat_summary <- flights[, .(
  Records      = .N,
  Departures   = sum(DepPerformed),
  Pct_Deps     = round(sum(DepPerformed) / sum(flights$DepPerformed) * 100, 1)
), by = Category][order(-Departures)]

cat("\n=== CATEGORY BREAKDOWN ===\n")
print(cat_summary)
fwrite(cat_summary, file.path(TBL_DIR, "category_breakdown.csv"))

# --- 4. Stratum summary -----
stratum_summary <- flights[, .(
  N_h          = .N,
  Deps         = sum(DepPerformed),
  Widebody     = sum(is_widebody),
  Narrowbody   = sum(Category == "Narrowbody"),

```

```

    Regional      = sum(Category == "Regional"),
    p_h           = round(mean(is_widebody), 4),
    S_h           = round(sd(is_widebody), 4)
  ), by = dist_band][order(dist_band)]

# Add stratum weight ( $W_h = N_h / N$ )
stratum_summary[, W_h := round(N_h / sum(N_h), 4)]

cat("\n=== STRATUM SUMMARY (population parameters) ===\n")
print(stratum_summary)
fwrite(stratum_summary, file.path(TBL_DIR, "stratum_summary.csv"))

# --- 5. Top competing routes -----
# Routes where BOTH narrowbody and widebody aircraft are used
route_mix <- flights[, .(
  Deps      = sum(DepPerformed),
  N_Wide    = sum(is_widebody),
  N_Narrow  = sum(Category == "Narrowbody"),
  Pct_Wide  = round(mean(is_widebody) * 100, 1),
  Carriers  = uniqueN(Carrier),
  Types     = paste(sort(unique(AircraftName)), collapse = "; ")
), by = .(OriginAirport, DestAirport, Distance)]

# Filter to routes with both categories and meaningful volume
competing <- route_mix[N_Wide > 0 & N_Narrow > 0 & Deps >= 500]
competing <- competing[order(-Deps)][1:20]

cat("\n=== TOP 20 COMPETING ROUTES (both narrow + widebody, 500 deps) ===\n")
print(competing[, .(OriginAirport, DestAirport, Distance, Deps,
  Pct_Wide, Carriers)])
fwrite(competing, file.path(TBL_DIR, "competing_routes.csv"))

# --- 6. Fleet mix by carrier on long-haul routes -----
# Which carriers use widebodies vs narrowbodies on 2000+ mi routes
long_haul <- flights[Distance >= 2000]
carrier_mix <- long_haul[, .(
  Deps      = sum(DepPerformed),
  Pct_Wide  = round(sum(is_widebody) / .N * 100, 1),
  Top_Types = paste(head(sort(unique(AircraftName))), 5), collapse = "; ")
), by = Carrier][Deps >= 1000][order(-Deps)]

cat("\n=== CARRIER FLEET MIX ON LONG-HAUL (2000 mi) ===\n")
print(carrier_mix)
fwrite(carrier_mix, file.path(TBL_DIR, "carrier_longhaul_mix.csv"))

```

```

# --- 7. Plots -----

# Plot 1: Widebody proportion by distance band
p1 <- ggplot(stratum_summary, aes(x = dist_band, y = p_h)) +
  geom_col(fill = "#2C5F8A", width = 0.7) +
  geom_text(aes(label = paste0(round(p_h * 100, 1), "%"),
    vjust = -0.5, size = 3.5) +
  scale_y_continuous(labels = percent_format(), expand = expansion(mult = c(0, 0.15))) +
  labs(
    title = "Widebody Proportion by Route Distance",
    subtitle = "U.S. domestic scheduled passenger jets, Jan-Dec 2025",
    x = "Nonstop distance (miles)",
    y = "Proportion of segment records using widebody aircraft"
  ) +
  theme_minimal(base_size = 12) +
  theme(panel.grid.major.x = element_blank())

ggsave(file.path(FIG_DIR, "widebody_pct_by_distance.png"), p1,
  width = 8, height = 5, dpi = 300)
cat("\nSaved: widebody_pct_by_distance.png\n")

# Plot 2: Stacked bar - category mix by distance band
mix_long <- flights[, .(Count = .N), by = .(dist_band, Category)]
mix_long[, Pct := Count / sum(Count), by = dist_band]
mix_long[, Category := factor(Category, levels = c("Regional", "Narrowbody",
  ↪ "Widebody"))]

p2 <- ggplot(mix_long, aes(x = dist_band, y = Pct, fill = Category)) +
  geom_col(width = 0.7) +
  scale_y_continuous(labels = percent_format()) +
  scale_fill_manual(values = c("Regional" = "#A8D5BA", "Narrowbody" = "#2C5F8A",
    "Widebody" = "#D94F3D")) +
  labs(
    title = "Fleet Category Mix by Route Distance",
    subtitle = "U.S. domestic scheduled passenger jets, Jan-Dec 2025",
    x = "Nonstop distance (miles)",
    y = "Proportion of segment records",
    fill = "Aircraft\ncategory"
  ) +
  theme_minimal(base_size = 12) +
  theme(panel.grid.major.x = element_blank())

ggsave(file.path(FIG_DIR, "fleet_mix_by_distance.png"), p2,

```

```

        width = 8, height = 5, dpi = 300)
cat("Saved: fleet_mix_by_distance.png\n")

# Plot 3: Stratum sizes
p3 <- ggplot(stratum_summary, aes(x = dist_band, y = N_h)) +
  geom_col(fill = "#5B8FA8", width = 0.7) +
  geom_text(aes(label = format(N_h, big.mark = ",")),
            vjust = -0.5, size = 3.2) +
  scale_y_continuous(labels = comma_format(), expand = expansion(mult = c(0, 0.12))) +
  labs(
    title = "Population Size by Distance Band (Stratum)",
    subtitle = "Short-haul segments vastly outnumber long-haul - motivates
      ↪ stratification",
    x = "Nonstop distance (miles)",
    y = "Number of segment records (N_h)"
  ) +
  theme_minimal(base_size = 12) +
  theme(panel.grid.major.x = element_blank())

ggsave(file.path(FIG_DIR, "stratum_sizes.png"), p3,
        width = 8, height = 5, dpi = 300)
cat("Saved: stratum_sizes.png\n")

cat("\n02_eda.R completed.\n")

```

7.4 Stratified Sampling

```

# =====
# 03_sample.R
# DATA 407 Project - Narrowbody vs Widebody Fleet Deployment
# Ronit Buti (80223837)
#
# Purpose: Implement stratified random sampling of widebody proportion.
#          Compute proportional and Neyman (optimal) allocation, draw samples,
#          and save for analysis in 04_analysis.R.
#
# Input:   data/processed/flights_clean.csv
# Output:  output/tables/allocation_table.csv
#          data/processed/sample_neyman.csv
#          data/processed/sample_proportional.csv
#          data/processed/sample_srs.csv
# =====

source("R/00_setup.R")

```

```

# --- 1. Load cleaned data -----
cat("Loading cleaned data...\n")
flights <- fread(CLEAN_FILE)
flights[, dist_band := factor(dist_band, levels = DIST_LABELS)]

# Drop rows with missing distance band (Distance = 0 artifacts)
flights <- flights[!is.na(dist_band)]

N <- nrow(flights)
n <- TOTAL_SAMPLE_SIZE

cat(" Population size N =", format(N, big.mark = ","), "\n")
cat(" Total sample size n =", n, "\n\n")

# --- 2. Population parameters by stratum -----
# N_h = population size in stratum h
# p_h = population proportion (widebody) in stratum h
# S_h = population standard deviation in stratum h
# W_h = N_h / N = stratum weight

pop_params <- flights[, .(
  N_h = .N,
  p_h = mean(is_widebody),
  S_h = sd(is_widebody)
), by = dist_band][order(dist_band)]

pop_params[, W_h := N_h / N]

cat("=== POPULATION PARAMETERS BY STRATUM ===\n")
print(pop_params)

# Overall population proportion (true value we're estimating)
p_true <- mean(flights$is_widebody)
cat("\nTrue population proportion (widebody): p =", round(p_true, 4), "\n\n")

# --- 3. Compute allocations -----

# Proportional allocation: n_h = n * (N_h / N)
pop_params[, n_prop := round(n * W_h)]

# Adjust rounding so total = n exactly
diff_prop <- n - sum(pop_params$n_prop)
if (diff_prop != 0) {

```

```

  idx_max <- which.max(pop_params$N_h)
  pop_params[idx_max, n_prop := n_prop + diff_prop]
}

# Neyman (optimal) allocation:  $n_h = N_h * S_h$ 
pop_params[, NhSh := N_h * S_h]
pop_params[, n_neyman := round(n * NhSh / sum(NhSh))]

# Adjust rounding
diff_ney <- n - sum(pop_params$n_neyman)
if (diff_ney != 0) {
  idx_max <- which.max(pop_params$NhSh)
  pop_params[idx_max, n_neyman := n_neyman + diff_ney]
}

# Ensure no stratum is allocated more than its population
pop_params[, n_prop := pmin(n_prop, N_h)]
pop_params[, n_neyman := pmin(n_neyman, N_h)]

# Display allocation comparison
alloc_table <- pop_params[, .(dist_band, N_h, W_h, p_h, S_h, n_prop, n_neyman)]
cat("=== ALLOCATION TABLE ===\n")
print(alloc_table)
cat("\nTotal proportional:", sum(alloc_table$n_prop), "\n")
cat("Total Neyman:", sum(alloc_table$n_neyman), "\n")

fwrite(alloc_table, file.path(TBL_DIR, "allocation_table.csv"))

# --- 4. Draw stratified samples -----
cat("\nDrawing samples (seed = 654)...\n")
set.seed(654)

# Helper: draw stratified sample given allocation vector
draw_stratified <- function(data, params, alloc_col) {
  samples <- list()
  for (i in seq_len(nrow(params))) {
    band <- params$dist_band[i]
    nh <- params[[alloc_col]][i]
    if (nh == 0) next
    stratum_data <- data[dist_band == band]
    idx <- sample(nrow(stratum_data), size = nh, replace = FALSE)
    samples[[length(samples) + 1]] <- stratum_data[idx]
  }
  rbindlist(samples)
}

```

```

}

# Draw stratified samples
sample_prop <- draw_stratified(flights, pop_params, "n_prop")
sample_neyman <- draw_stratified(flights, pop_params, "n_neyman")

# Draw SRS for comparison
sample_srs <- flights[sample(N, size = n, replace = FALSE)]

cat(" Proportional sample: n =", nrow(sample_prop), "\n")
cat(" Neyman sample:      n =", nrow(sample_neyman), "\n")
cat(" SRS sample:         n =", nrow(sample_srs), "\n")

# --- 5. Quick sample verification -----
cat("\n=== SAMPLE COMPOSITION CHECK ===\n")

cat("\nProportional sample by stratum:\n")
print(sample_prop[, .(N, widebody = sum(is_widebody),
  p_hat = round(mean(is_widebody), 4)), by =
  ↪ dist_band][order(dist_band)])

cat("\nNeyman sample by stratum:\n")
print(sample_neyman[, .(N, widebody = sum(is_widebody),
  p_hat = round(mean(is_widebody), 4)), by =
  ↪ dist_band][order(dist_band)])

cat("\nSRS sample by stratum (uncontrolled):\n")
print(sample_srs[, .(N, widebody = sum(is_widebody),
  p_hat = round(mean(is_widebody), 4)), by =
  ↪ dist_band][order(dist_band)])

# --- 6. Save samples -----
fwrite(sample_prop, "data/processed/sample_proportional.csv")
fwrite(sample_neyman, "data/processed/sample_neyman.csv")
fwrite(sample_srs, "data/processed/sample_srs.csv")

cat("\nSamples saved to data/processed/.\n")
cat("03_sample.R completed.\n")

```

7.5 Analysis and Simulation

```

# =====
# 04_analysis.R
# DATA 407 Project - Narrowbody vs Widebody Fleet Deployment

```

```

# Ronit Buti (80223837)
#
# Purpose: Compute stratified estimates of widebody proportion with CIs,
#          compare SRS vs proportional vs Neyman allocation, and run a
#          simulation to visualize the sampling distribution.
#
# Input:   data/processed/flights_clean.csv
#          data/processed/sample_proportional.csv
#          data/processed/sample_neyman.csv
#          data/processed/sample_srs.csv
#          output/tables/allocation_table.csv
# Output:  output/tables/estimation_comparison.csv
#          output/tables/stratum_estimates_neyman.csv
#          output/tables/simulation_results.csv
# =====

source("R/00_setup.R")

# --- 1. Load data -----
cat("Loading population and samples...\n")
flights      <- fread(CLEAN_FILE)
sample_prop  <- fread("data/processed/sample_proportional.csv")
sample_ney   <- fread("data/processed/sample_neyman.csv")
sample_srs   <- fread("data/processed/sample_srs.csv")
alloc_table  <- fread(file.path(TBL_DIR, "allocation_table.csv"))

flights[, dist_band := factor(dist_band, levels = DIST_LABELS)]
sample_prop[, dist_band := factor(dist_band, levels = DIST_LABELS)]
sample_ney[, dist_band := factor(dist_band, levels = DIST_LABELS)]
sample_srs[, dist_band := factor(dist_band, levels = DIST_LABELS)]

N      <- nrow(flights)
n      <- TOTAL_SAMPLE_SIZE
p_true <- mean(flights$is_widebody)

cat(" True population proportion p =", round(p_true, 4), "\n\n")

# --- 2. Estimator functions -----

# Stratified estimator:  $p_{\text{hat\_str}} = \sum W_h * p_{\text{hat\_h}}$ 
# Variance:  $\text{Var}(p_{\text{hat\_str}}) = \sum (1 - n_h/N_h) * W_h^2 * s_h^2 / n_h$ 
stratified_estimate <- function(sample_dt, alloc_dt) {
  H <- nrow(alloc_dt)
  p_str <- 0

```

```

var_str <- 0
stratum_results <- list()

for (i in seq_len(H)) {
  band <- alloc_dt$dist_band[i]
  N_h <- alloc_dt$N_h[i]
  W_h <- alloc_dt$W_h[i]

  sub <- sample_dt[dist_band == band]
  n_h <- nrow(sub)
  p_hat_h <- mean(sub$is_widebody)
  s2_h <- var(sub$is_widebody) # sample variance (denominator n_h - 1)

  p_str <- p_str + W_h * p_hat_h
  var_str <- var_str + (1 - n_h / N_h) * W_h^2 * s2_h / n_h

  stratum_results[[i]] <- data.table(
    dist_band = band, N_h = N_h, n_h = n_h, W_h = W_h,
    p_hat_h = round(p_hat_h, 4), s2_h = round(s2_h, 6),
    ci_lo_h = round(p_hat_h - 1.96 * sqrt((1 - n_h/N_h) * s2_h / n_h), 4),
    ci_hi_h = round(p_hat_h + 1.96 * sqrt((1 - n_h/N_h) * s2_h / n_h), 4)
  )
}

se_str <- sqrt(var_str)
list(
  estimate = p_str,
  se = se_str,
  ci_lo = p_str - 1.96 * se_str,
  ci_hi = p_str + 1.96 * se_str,
  strata = rbindlist(stratum_results)
)
}

# SRS estimator: p_hat = y-bar, Var = (1 - n/N) * s^2 / n
srs_estimate <- function(sample_dt) {
  p_hat <- mean(sample_dt$is_widebody)
  s2 <- var(sample_dt$is_widebody)
  se <- sqrt((1 - n / N) * s2 / n)
  list(
    estimate = p_hat,
    se = se,
    ci_lo = p_hat - 1.96 * se,
    ci_hi = p_hat + 1.96 * se
  )
}

```

```

)
}

# --- 3. Compute estimates -----
cat("=== ESTIMATION RESULTS ===\n\n")

res_srs <- srs_estimate(sample_srs)
res_prop <- stratified_estimate(sample_prop, alloc_table)
res_ney <- stratified_estimate(sample_ney, alloc_table)

# Comparison table
comparison <- data.table(
  Method = c("SRS", "Stratified (proportional)", "Stratified (Neyman)"),
  Estimate = round(c(res_srs$estimate, res_prop$estimate, res_ney$estimate), 4),
  SE = round(c(res_srs$se, res_prop$se, res_ney$se), 4),
  CI_lower = round(c(res_srs$ci_lo, res_prop$ci_lo, res_ney$ci_lo), 4),
  CI_upper = round(c(res_srs$ci_hi, res_prop$ci_hi, res_ney$ci_hi), 4),
  CI_width = round(c(res_srs$ci_hi - res_srs$ci_lo,
                    res_prop$ci_hi - res_prop$ci_lo,
                    res_ney$ci_hi - res_ney$ci_lo), 4),
  Variance = c(res_srs$se^2, res_prop$se^2, res_ney$se^2)
)
comparison[, Var_pct_of_SRS := round(Variance / Variance[1] * 100, 1)]

cat("True proportion:", round(p_true, 4), "\n\n")
print(comparison)

fwrite(comparison, file.path(TBL_DIR, "estimation_comparison.csv"))

# Stratum-level detail for Neyman
cat("\n=== STRATUM-LEVEL ESTIMATES (Neyman) ===\n")
print(res_ney$strata)
fwrite(res_ney$strata, file.path(TBL_DIR, "stratum_estimates_neyman.csv"))

# --- 4. Simulation: repeated sampling -----
# Draw B stratified samples (Neyman) and B SRS samples to compare the
# sampling distributions of the two estimators.

B <- 2000
cat("\n=== SIMULATION: B =", B, "replications ===\n")

sim_srs <- numeric(B)
sim_prop <- numeric(B)
sim_neyman <- numeric(B)

```

```

for (b in seq_len(B)) {
  # SRS
  idx_srs <- sample(N, n, replace = FALSE)
  sim_srs[b] <- mean(flights$is_widebody[idx_srs])

  # Stratified (proportional)
  samp_p <- list()
  for (i in seq_len(nrow(alloc_table))) {
    band <- alloc_table$dist_band[i]
    nh <- alloc_table$n_prop[i]
    pool <- which(flights$dist_band == band)
    samp_p[[i]] <- pool[sample(length(pool), nh)]
  }
  idx_prop <- unlist(samp_p)
  # Weighted estimate
  for (i in seq_len(nrow(alloc_table))) {
    band <- alloc_table$dist_band[i]
    nh <- alloc_table$n_prop[i]
  }
  # Compute stratified mean
  p_b <- 0
  for (i in seq_len(nrow(alloc_table))) {
    band <- alloc_table$dist_band[i]
    W_h <- alloc_table$W_h[i]
    sub_idx <- idx_prop[flights$dist_band[idx_prop] == band]
    p_b <- p_b + W_h * mean(flights$is_widebody[sub_idx])
  }
  sim_prop[b] <- p_b

  # Stratified (Neyman)
  samp_n <- list()
  for (i in seq_len(nrow(alloc_table))) {
    band <- alloc_table$dist_band[i]
    nh <- alloc_table$n_neyman[i]
    pool <- which(flights$dist_band == band)
    samp_n[[i]] <- pool[sample(length(pool), nh)]
  }
  idx_ney <- unlist(samp_n)
  p_b <- 0
  for (i in seq_len(nrow(alloc_table))) {
    band <- alloc_table$dist_band[i]
    W_h <- alloc_table$W_h[i]
    sub_idx <- idx_ney[flights$dist_band[idx_ney] == band]

```

```

    p_b <- p_b + W_h * mean(flights$sis_widebody[sub_idx])
  }
  sim_neyman[b] <- p_b
}

sim_results <- data.table(
  Method = c("SRS", "Stratified (proportional)", "Stratified (Neyman)"),
  Mean   = round(c(mean(sim_srs), mean(sim_prop), mean(sim_neyman)), 5),
  SD     = round(c(sd(sim_srs), sd(sim_prop), sd(sim_neyman)), 5),
  MSE    = round(c(mean((sim_srs - p_true)^2),
                    mean((sim_prop - p_true)^2),
                    mean((sim_neyman - p_true)^2)), 8)
)
sim_results[, Relative_SD := round(SD / SD[1] * 100, 1)]

cat("\nSimulation summary (B = ", B, "):\n", sep = "")
cat("True p =", round(p_true, 4), "\n\n")
print(sim_results)

# Save simulation draws for plotting in 05_visualize.R
sim_draws <- data.table(
  SRS = sim_srs, Proportional = sim_prop, Neyman = sim_neyman
)
fwrite(sim_draws, file.path(TBL_DIR, "simulation_draws.csv"))
fwrite(sim_results, file.path(TBL_DIR, "simulation_results.csv"))

cat("\n04_analysis.R completed.\n")

```

7.6 Visualization

```

# =====
# 05_visualize.R
# DATA 407 Project - Narrowbody vs Widebody Fleet Deployment
# Ronit Buti (80223837)
#
# Purpose: Generate figures for the report.
#         EDA plots are in 02_eda.R; this script focuses on sampling
#         results and method comparisons.
#
# Input:  output/tables/estimation_comparison.csv
#         output/tables/allocation_table.csv
#         output/tables/simulation_draws.csv
#         output/tables/stratum_estimates_neyman.csv
#         data/processed/flights_clean.csv

```

```

# Output:  output/figures/ci_comparison.png
#          output/figures/allocation_comparison.png
#          output/figures/sampling_distribution.png
#          output/figures/route_competition.png
# =====

source("R/00_setup.R")

# --- Load results -----
comparison <- fread(file.path(TBL_DIR, "estimation_comparison.csv"))
alloc_table <- fread(file.path(TBL_DIR, "allocation_table.csv"))
sim_draws <- fread(file.path(TBL_DIR, "simulation_draws.csv"))
strat_est <- fread(file.path(TBL_DIR, "stratum_estimates_neyman.csv"))
flights <- fread(CLEAN_FILE)
flights[, dist_band := factor(dist_band, levels = DIST_LABELS)]

p_true <- mean(flights$is_widebody)

# --- Figure 1: CI comparison -----
# Side-by-side confidence intervals for SRS, proportional, and Neyman
comparison[, Method := factor(Method,
  levels = c("SRS", "Stratified (proportional)", "Stratified (Neyman)"))]

p_ci <- ggplot(comparison, aes(x = Method, y = Estimate)) +
  geom_hline(yintercept = p_true, linetype = "dashed", color = "#888888", linewidth =
    ↪ 0.6) +
  geom_pointrange(aes(ymin = CI_lower, ymax = CI_upper),
    size = 0.8, linewidth = 1, color = "#2C5F8A") +
  geom_text(aes(label = paste0("Width: ", sprintf("%.4f", CI_width))),
    hjust = -0.2, size = 3.2) +
  annotate("text", x = 0.55, y = p_true, label = paste("True p =", round(p_true, 4)),
    hjust = 0, vjust = -0.8, size = 3.2, color = "#888888") +
  coord_flip() +
  scale_y_continuous(labels = scales::number_format(accuracy = 0.001)) +
  labs(
    title = "95% Confidence Intervals by Sampling Method",
    subtitle = "n = 1,000 | Dashed line = true population proportion",
    x = NULL, y = "Estimated widebody proportion"
  ) +
  theme_minimal(base_size = 12) +
  theme(panel.grid.major.y = element_blank())

ggsave(file.path(FIG_DIR, "ci_comparison.png"), p_ci,
  width = 9, height = 4, dpi = 300)

```

```

cat("Saved: ci_comparison.png\n")

# --- Figure 2: Allocation comparison -----
# How proportional vs Neyman distributes samples across strata
alloc_long <- melt(alloc_table,
  id.vars = "dist_band",
  measure.vars = c("n_prop", "n_neyman"),
  variable.name = "Allocation",
  value.name = "n_h"
)
alloc_long[, Allocation := fifelse(Allocation == "n_prop", "Proportional", "Neyman")]
alloc_long[, dist_band := factor(dist_band, levels = DIST_LABELS)]

p_alloc <- ggplot(alloc_long, aes(x = dist_band, y = n_h, fill = Allocation)) +
  geom_col(position = "dodge", width = 0.65) +
  geom_text(aes(label = n_h), position = position_dodge(width = 0.65),
    vjust = -0.4, size = 3) +
  scale_fill_manual(values = c("Proportional" = "#A8D5BA", "Neyman" = "#2C5F8A")) +
  scale_y_continuous(expand = expansion(mult = c(0, 0.15))) +
  labs(
    title = "Sample Allocation by Distance Band",
    subtitle = "Neyman shifts sampling effort toward high-variance long-haul strata",
    x = "Nonstop distance (miles)", y = "Sample size (n_h)",
    fill = "Allocation"
  ) +
  theme_minimal(base_size = 12) +
  theme(panel.grid.major.x = element_blank())

ggsave(file.path(FIG_DIR, "allocation_comparison.png"), p_alloc,
  width = 8, height = 5, dpi = 300)
cat("Saved: allocation_comparison.png\n")

# --- Figure 3: Sampling distribution (simulation) -----
# Overlay density curves from repeated sampling
sim_long <- melt(sim_draws,
  measure.vars = c("SRS", "Neyman"),
  variable.name = "Method",
  value.name = "Estimate"
)

p_sim <- ggplot(sim_long, aes(x = Estimate, fill = Method, color = Method)) +
  geom_density(alpha = 0.3, linewidth = 0.8) +
  geom_vline(xintercept = p_true, linetype = "dashed", color = "#333333", linewidth =
    ↪ 0.5) +

```

```

  annotate("text", x = p_true, y = Inf, label = paste("True p =", round(p_true, 4)),
    hjust = -0.1, vjust = 2, size = 3.2) +
  scale_fill_manual(values = c("SRS" = "#D94F3D", "Neyman" = "#2C5F8A")) +
  scale_color_manual(values = c("SRS" = "#D94F3D", "Neyman" = "#2C5F8A")) +
  labs(
    title = "Sampling Distribution: SRS vs Stratified (Neyman)",
    subtitle = paste("B = 2,000 replications | n = 1,000 each"),
    x = "Estimated widebody proportion",
    y = "Density"
  ) +
  theme_minimal(base_size = 12)

ggsave(file.path(FIG_DIR, "sampling_distribution.png"), p_sim,
  width = 8, height = 5, dpi = 300)
cat("Saved: sampling_distribution.png\n")

# --- Figure 4: Route-level competition (LAX-JFK example) -----
# Aircraft types used on a single transcontinental route
laxjfk <- flights[
  (OriginAirport == "LAX" & DestAirport == "JFK") |
  (OriginAirport == "JFK" & DestAirport == "LAX")
]

type_summary <- laxjfk[, .(
  Deps = sum(DepPerformed)
), by = .(AircraftName, Category)][order(-Deps)]

# Keep top types for readability
type_summary <- type_summary[Deps >= 100]
type_summary[, AircraftName := factor(AircraftName, levels = rev(AircraftName))]

p_route <- ggplot(type_summary, aes(x = AircraftName, y = Deps, fill = Category)) +
  geom_col(width = 0.7) +
  coord_flip() +
  scale_fill_manual(values = c("Narrowbody" = "#2C5F8A", "Widebody" = "#D94F3D",
    "Regional" = "#A8D5BA")) +
  scale_y_continuous(labels = comma_format()) +
  labs(
    title = "JFK LAX: Aircraft Types Competing on the Same Route",
    subtitle = "2,475 miles - narrowbodies and widebodies in direct competition",
    x = NULL, y = "Total departures (2025)",
    fill = "Category"
  ) +
  theme_minimal(base_size = 12) +

```

```

theme(panel.grid.major.y = element_blank())

ggsave(file.path(FIG_DIR, "route_competition.png"), p_route,
        width = 9, height = 5, dpi = 300)
cat("Saved: route_competition.png\n")

# --- Figure 5: Stratum-level estimates with CIs (Neyman) -----
strat_est[, dist_band := factor(dist_band, levels = DIST_LABELS)]

# Get true population proportions for comparison
pop_p <- flights[, .(p_true_h = mean(is_widebody)), by = dist_band]
strat_est <- merge(strat_est, pop_p, by = "dist_band")

p_strata <- ggplot(strat_est, aes(x = dist_band)) +
  geom_pointrange(aes(y = p_hat_h, ymin = ci_lo_h, ymax = ci_hi_h),
                 color = "#2C5F8A", size = 0.7, linewidth = 0.9) +
  geom_point(aes(y = p_true_h), shape = 4, size = 3, color = "#D94F3D", stroke = 1.2) +
  scale_y_continuous(labels = percent_format()) +
  labs(
    title = "Stratum-Level Widebody Proportion Estimates (Neyman Allocation)",
    subtitle = "Blue = sample estimate with 95% CI | Red x = true population value",
    x = "Nonstop distance (miles)",
    y = "Widebody proportion"
  ) +
  theme_minimal(base_size = 12) +
  theme(panel.grid.major.x = element_blank())

ggsave(file.path(FIG_DIR, "stratum_estimates.png"), p_strata,
        width = 8, height = 5, dpi = 300)
cat("Saved: stratum_estimates.png\n")

cat("\n05_visualize.R completed. All figures saved to", FIG_DIR, "\n")

```