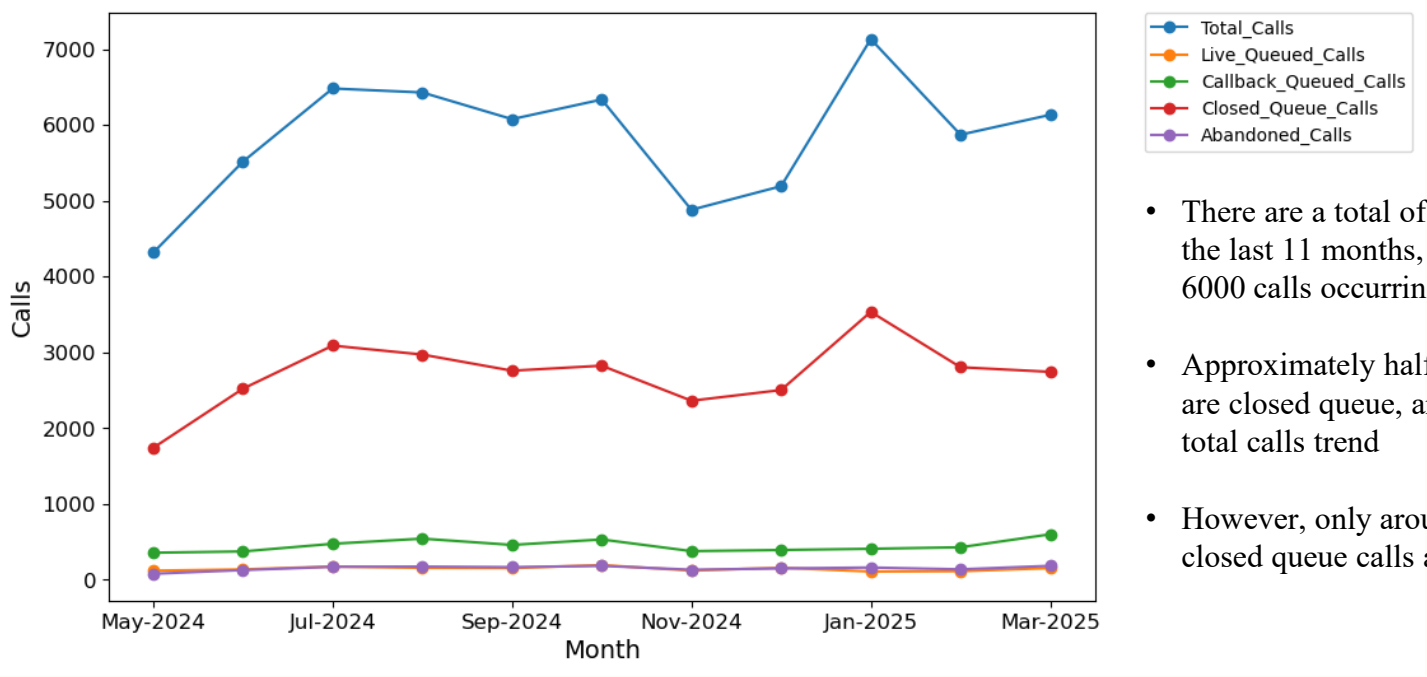
Legal Aid Chicago: Data summary and integrity

Mia, Lila, Maia

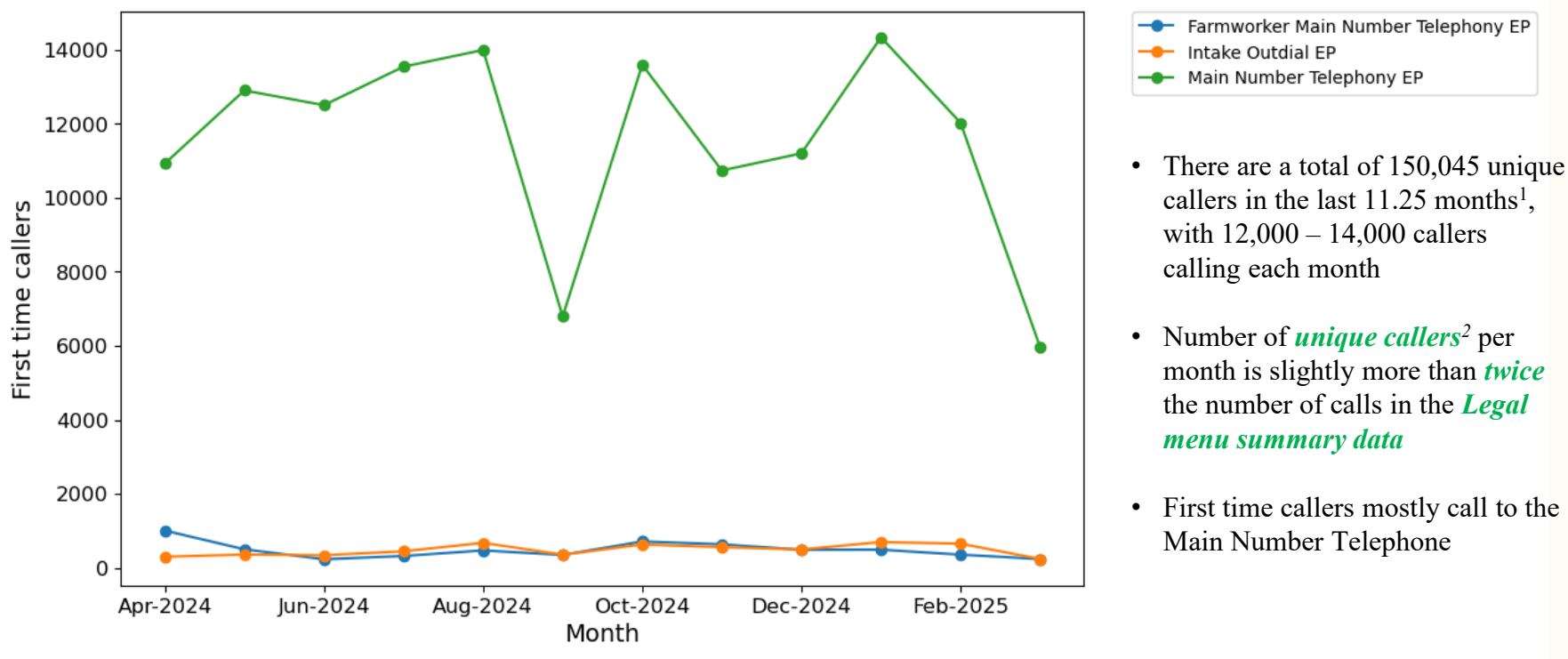
Table of Contents

Legal Menu Summary Reports (LMSR)



- There are a total of 64,349 calls in the last 11 months, with 4000 – 6000 calls occurring each month
- Approximately half the total calls are closed queue, and mirror the total calls trend
- However, only around 20% of the closed queue calls are called back

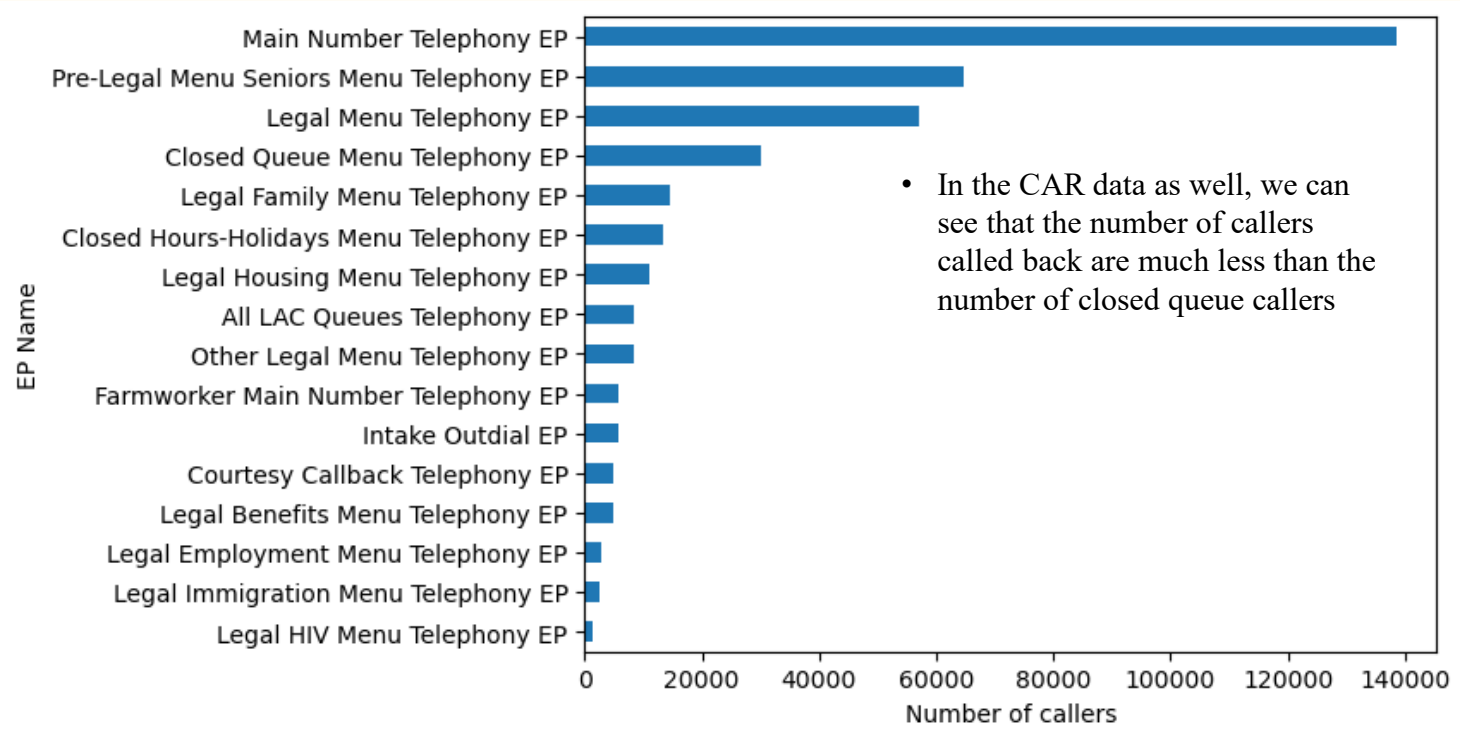
CAR Data – First time entry points (EPs)



- There are a total of 150,045 unique callers in the last 11.25 months[1], with 12,000 – 14,000 callers calling each month
- Number of ***unique callers***[2] per month is slightly more than ***twice*** the number of calls in the ***Legal menu summary data***
- First time callers mostly call to the Main Number Telephone

Note: 1. Data is from 7th April 2024 to 15th March 2025
2. Number of unique callers are given by the number of distinct values of 'Contact Session ID'

CAR Data – all entry points (EPs)



- In the CAR data as well, we can see that the number of callers called back are much less than the number of closed queue callers

Note: 1. Data is from 7th April 2024 to 15th March 2025
2. Number of unique callers are given by the number of distinct values of 'Contact Session ID'

CAR Data

Contact Session ID	EP Name	Flow Name	Activity Name	Activity Start Timestamp	Queue Name	Agent Name	Hour
94c8fb41-8baa-4aed-bd42-9150c02ade3a	Closed Hours-Holidays Menu Telephony EP	NaN	NaN	2024-12-29 09:49:05	NaN	NaN	9
94c8fb41-8baa-4aed-bd42-9150c02ade3a	Closed Hours-Holidays Menu Telephony EP	NaN	NaN	2024-12-29 09:49:05	NaN	NaN	9
d139ea26-46bf-4de2-91eb-378d6ce1411c	Closed Hours-Holidays Menu Telephony EP	NaN	NaN	2024-12-29 11:01:43	NaN	NaN	11
d139ea26-46bf-4de2-91eb-378d6ce1411c	Closed Hours-Holidays Menu Telephony EP	NaN	NaN	2024-12-29 11:01:43	NaN	NaN	11

- There are 33,194 ***duplicate observations*** in the data, which we will ***discard***

Question

- As the number of unique callers in the CAR data (around 150k) are more than twice the number of calls (around 64k) in the Legal Menu Summaries Data, which calls of the CAR data might not be counted in the Legal Menu summaries data?

Question answered

Q) Can you figure out how callers are getting to the front desk when they are not given that as a menu option – how to know that?

A) In addition to getting to the front desk via the main menu, callers are getting to the front desk via the Legal Menu after they have been to:

1. Pre-Legal Menu Seniors Menu – can press ‘2’ or nothing to go to the Legal Menu
2. Other Legal Menu – can press ‘9’ or nothing or ‘invalid input’ to go to the Legal Menu
3. Legal Housing Menu - can press ‘9’ or nothing or ‘invalid input’ to go to the Legal Menu
4. Legal Family Menu – can press ‘9’ or nothing or ‘invalid input’ to go to the Legal Menu
5. Legal HIV Menu – can press nothing or ‘invalid input’ to go to the Legal Menu

Most of them (around 1600 in a year) are going through the Pre-Legal Menu Seniors Menu.

Question answered (cont.)

Q) Can you figure out how callers are getting to the front desk when they are not given that as a menu option – how to know that?

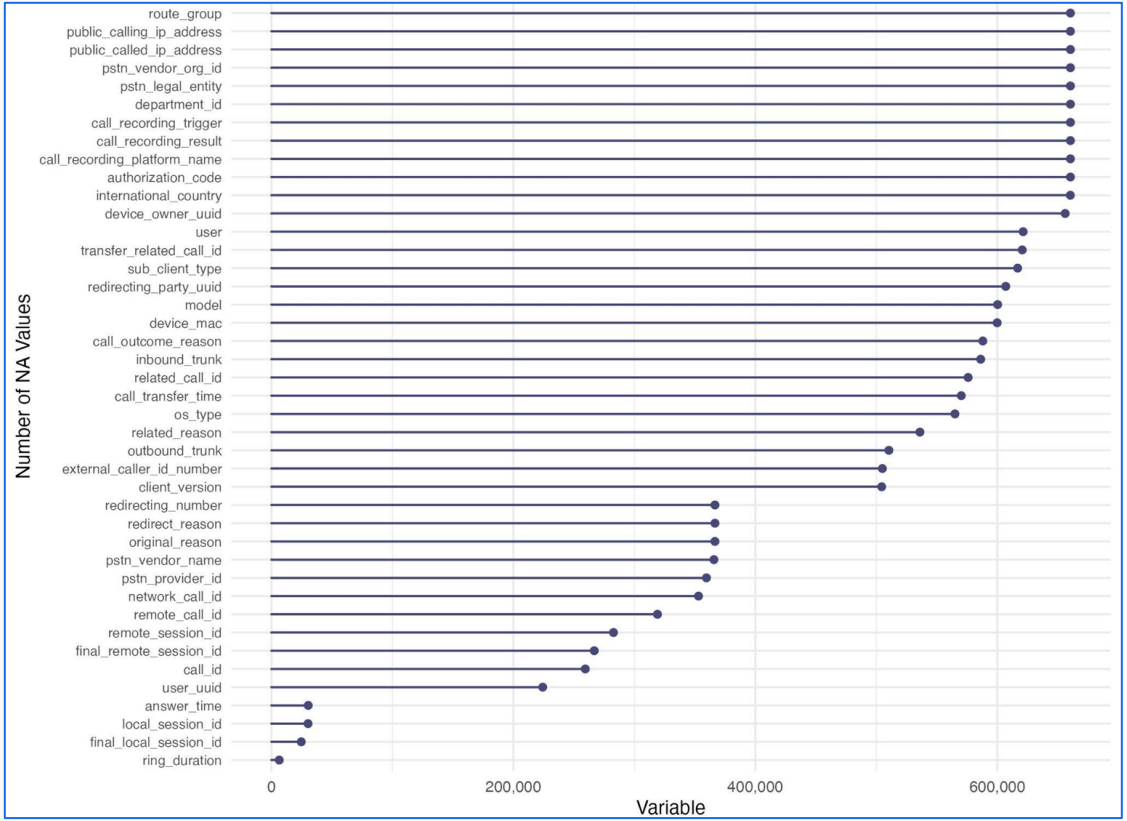
- Out of 150k distinct callers, only 11k reach the front desk
- Approximately 9.5k callers reach the front desk via the Main number, and 1.5k reach via the Legal Menu
- Out of the 11k callers, 9.3k reach the Front desk via the Main number
- Among the remaining 1.7k, 1.6k are going to the Front desk by pressing ‘2’ or nothing in the Pre-Legal Menu Seniors Menu

All Calls Data Overview

- There are **660,413 rows** and **65 columns** in the merged data
- However, only March 2025 contains all 65 columns
  - July, August, September, October, and December 2024 contain 56, missing the following:
    - User
    - External caller ID number
    - Device owner UUID
    - Call recording platform name
    - Call recording result
    - Call recording trigger
    - Redirecting party UUID
    - Public calling IP address
    - Public called IP address
  - April, May, and June 2024, along with January and February 2025 have 64 columns, missing user

All Calls Missing Data

- 10 variables are entirely “NA”
- However, some NA values simply show up if a variable is not relevant for a particular call
- For example, public calling IP address is only applicable to calls made in India locations



*Variables not included in this plot have no NA values

All Calls Data - Closer Look at NA Values

Most of the occurrences of NA values can be easily explained

- *international_country*: only displayed for international calls
- *pstn_legal_entity*: exclusively for Cisco Calling Plans
  - *pstn_vendor_org_id*: displays the Cisco Calling plan's org UUID
- *call_recording_platform, call_recording_trigger*, *call_recording_result*: only applicable if call was recorded
- *public_calling_ip_address*, *public_called_ip_address*: only relevant for India locations
- *route_group*: if present, this field's only reported in originating (outbound) records
- *authorization_code*: only present when an authorization code is assigned to a location or site for outgoing calls
- *device_owner_uuid*: only set when the device owner is different than the owner of the device who made or received the call

Other variables with high proportions of NA values:

- *department_id*: unique identifier for the user's department name
- *user*: represents the display name for the type of user who made or received the call

Full analysis of all variables with NA values (with corresponding counts and percentages) can be found in the Appendix

All Calls Variable Distributions

- The distribution of the two continuous variables are shown below
- Most categorical variables have thousands of levels; the total number of categories and example values for each non-missing variable can be found in the Appendix, as well as the supporting code

Variable	Mean	St. Dev.	Min	25th Pct	50th Pct	75th Pct	Max
duration	199.88	845.54	0	19	62	183	142257
ring_duration	5.00	8.24	0	0	0	8	243

Summary Statistics for Continuous Numeric Variables

Overlap Between Datasets

Overlap Across Datasets

- None of the IDs in All Calls appear in the Call Journey Data
- Both Call Journey Data and Legal Menu Summary Reports have queue information, but the Call Journey Data has a unique ID for each *caller* while LMSR shows the number of *calls*
  - These may contain overlap but need to determine a standardized way to identify the number of calls in CAR for each contact session ID

Questions & Clarifications

Clarifications

- Duplicate rows in the Call Journey Data
- Counting calls in the Call Journey Data
  - Using contact session ID and timestamp
  - Courtesy Callback Telephony EP
  - How is a “call” defined in the Legal Menu Summary Reports

Appendix

R Code LMSR - Data Cleaning

```r
#merging and cleaning data
file_list <- list.files(
  path = ".",
  pattern = "^(January|February|March|April|May|June|July|August|September|October|November|December)\\s\\d{4}\\.xlsx$",
  full.names = TRUE
)

read_and_clean <- function(file_path) {
  read_excel(file_path, skip = 2) %>%
    janitor::clean_names() %>%
    fill(legal_menu_option, menu_selection, queue_selection, group_suboption, .direction = "down") %>%
    mutate(month = tools::file_path_sans_ext(basename(file_path)))  #add month name for reference
}

all_months_cleaned <- map_dfr(file_list, read_and_clean)

View(all_months_cleaned)
write_csv(all_months_cleaned, "legal_menu_all_months_cleaned.csv")
```

R Code LMSR - Findings

```r
# Quantile Summary of Continuous Variables
df %>%
  select(total_calls, live_queued_calls, callback_queued_calls,
         closed_queue_calls, abandoned_calls) %>%
  summary() %>%
  kable()
```

```r
# Create frequency summary for categorical variables
categorical_summary <- df %>%
  select(where(is.character)) %>%
  map_df(~{
    tibble(
      top_5_values = paste(names(sort(table(.), decreasing = TRUE))[1:5], collapse = ", "),
      total_unique = n_distinct(.),
      num_missing = sum(is.na(.))
    )
  }, .id = "variable") %>%
  kable()

print(categorical_summary)
```

```r
#count for each month
df %>%
  count(month) %>%
  kable()
```

R Code All Calls - Data Cleaning

```r
months <- c(
  "April 2024", "May 2024", "June 2024", "July 2024", "August 2024", "September 2024",
  "October 2024", "December 2024", "January 2025", "February 2025", "March 2025"
  )

for (month in months) {
  var_name <- paste0(tolower(substr(month, 1, 3)), "_", substr(month, nchar(month) - 1, nchar(month)))
  path <- paste0("All Calls by Month/", month, ".xlsx")

  data <- read_excel(path) %>%
    clean_names() %>%
    mutate(across(everything(), as.character)) %>%
    mutate(across(everything(), ~ ifelse(. %in% c("na", "NA", ""), NA, .))) %>%
    mutate(month = month)

  assign(var_name, data)
  }

# combine months
all_calls <-
  bind_rows(apr_24, may_24, jun_24, jul_24, aug_24, sep_24, oct_24, dec_24, jan_25, feb_25, mar_25) %>%
  mutate(month = factor(
    month,
    levels = c("April 2024", "May 2024", "June 2024", "July 2024", "August 2024", "September 2024",
               "October 2024", "December 2024", "January 2025", "February 2025", "March 2025"), ordered = TRUE),
    across(where(is.logical), as.character),
    model = as.character(model),
    duration = as.numeric(duration),
    ring_duration = as.numeric(ring_duration),
    site_main_number = as.character(site_main_number),
    site_timezone = as.character(site_timezone),
    user_number = as.character(user_number)
  )

write_csv(all_calls, file = "all_calls.csv")
```

R Code All Calls - Data Overview, Missing Data, Month Breakdown

```r
# plot counts of missing data
all_calls %>%
  select(where(~ any(is.na(.)))) %>%
  naniar::gg_miss_var() +
  labs(x = "Number of Missing Values", y = "Variable") +
  scale_y_continuous(labels = scales::label_number(big.mark = ",")) +
  theme_minimal(base_size = 11) +
  theme(
    axis.text.y = element_text(size = 7),
    axis.text.x = element_text(size = 8),
    axis.title.x = element_text(size = 10),
    axis.title.y = element_text(size = 10)
  ) + coord_flip()
```

```r
# get skim of full data
skim_without_charts(all_calls)
```

```r
# find which columns are missing in some months
setdiff(names(mar_25), names(apr_24))
setdiff(names(mar_25), names(dec_24))

# get counts by month
total_calls <- all_calls %>% nrow()
month_breakdown <- all_calls %>%
  count(month) %>%
  mutate(Percent = round(n / total_calls * 100, 2)) %>%
  rename(Count = n, Month = month)

month_breakdown %>%
  kable()
```

R Code All Calls - Variable Distributions

```r
# function to get categorical vars summary
summarize_categoricals <- function(df, max_examples = 5) {
  # get names of logical, character, or factor variables
  cat_logical_vars <- names(df)[sapply(df, function(x) {(is.character(x) || is.factor(x) || is.logical(x)) && !all(is.na(x))})]

  summaries <- lapply(cat_logical_vars, function(var) {
    freq_table <- sort(table(df[[var]], useNA = "no"), decreasing = TRUE)
    total_unique <- length(freq_table)

    # skip if no non-missing values
    if (total_unique == 0) return(NULL)

    example_vals <- names(freq_table)[1:min(max_examples, total_unique)]
    example_counts <- as.numeric(freq_table[example_vals])
    example_str <- paste0(mapply(function(val, count) paste0(val, " (", count, ")"), example_vals, example_counts), collapse = "; ")

    if (total_unique > max_examples) {example_str <- paste0(example_str, "; ... [Total: ", total_unique, " categories]")}
    data.frame(variable = var, examples = example_str, total_categories = total_unique,stringsAsFactors = FALSE)
    })

  result <- do.call(rbind, summaries)

  if (!is.null(result)) {result}
  else {data.frame(variable = character(), examples = character(), total_categories = integer())}
  }

# get table
summary_table <- summarize_categoricals(all_calls)
knitr::kable(summary_table)
```

```r
# get numeric distributions
numeric_summary_tbl <- all_calls %>%
  select(where(is.numeric)) %>%
  summary()
numeric_summary_tbl %>% kable()
```

All Calls Data - Breakdown of NA Values

Variable	Number of NA Observations	Percent of NA Observations	Potential Explanation for NA Values
route_group	660413	100%	Only included for outbound calls
department_id	660413	100%	
authorization_code	660413	100%	Only present when an authorization code is assigned to a location or site for outgoing calls
pstn_legal_entity	660413	100%	Only present for calls made through Cisco Calling Plans
pstn_vendor_org_id	660413	100%	Only present for Cisco Calling Plans
call_recording_platform_name	660413	100%	Only present if a call recording platform is used for the call
call_recording_result	660413	100%	Only present when a call is recorded
call_recording_trigger	660413	100%	Only present when a call is recorded
public_calling_ip_address	660413	100%	Only used for calls made in India
public_called_ip_address	660413	100%	Only present for calls in India
international_country	660318	99.9%	Only displayed for international calls
device_owner_uuid	656095	99.3%	Only set when the device owner is different than the owner of the device who made/received the call
user	621318	94.1%	
transfer_related_call_id	620579	94%	Only relevant if call is involved in a transfer
sub_client_type	616782	93.4%	If the call is TO or FROM a mobile phone using Webex Go, the Client type display as SIP, and the Sub client type display as MOBILE_NETWORK
redirecting_party_uuid	606958	91.9%	A call needs to be redirected one or more times for this variable to have a value
model	600243	90.9%	
device_mac	599868	90.8%	The MAC address of the device, if known
call_outcome_reason	588004	89.0%	
inbound_trunk	586230	88.8%	Only present when a call is received from on-prem or PSTN via an inbound trunk under legacy call routing behavior
related_call_id	575851	87.2%	This field only gets populated if a new call is created as a result of some service activation

**Variables left blank are those with no inherent explanation behind NA values; may be truly missing*

All Calls Data - Breakdown of NA Values, cont.

Variable	Number of NA Observations	Percent of NA Observations	Potential Explanation for NA Values
call_transfer_time	570177	86.3%	Only if a transfer occurs
os_type	564960	85.5%	The operating system that the app is running on, if available
related_reason	536017	81.2%	Only present when a trigger leads to a change in call presence (e.g. redirected, no answer, etc.)
outbound_trunk	510373	77.3%	Only present when a call is routed through an outbound trunk to premises-based PSTN or on-prem deployment
external_caller_id	504955	76.5%	Only set when the control hub External caller ID phone number setting is either location number or other number from organization; not set when "Direct line/Ext" options has been selected
client_version	504415	76.4%	Only present when the call involves a Webex client application
original_reason	366584	55.5%	Only present when a call is redirected
redirect_reason	366584	55.5%	Only present when a call is redirected
redirecting_number	366584	55.5%	Only present when a call is redirected
pstn_vendor_name	365758	55.4%	Only present for calls involving PSTN services
pstn_provider_id	359576	54.4$	Only present for PSTN calls
network_call_id	353003	53.5%	
remote_call_id	319092	48.3%	
remote_session_id	282718	42.8%	
final_remote_session_id	266838	40.4%	
call_id	259413	39.3%	
user_uuid	224248	34%	
answer_time	30419	4.61%	Only for calls that get answered
local_session_id	30278	4.58%	
final_local_session_id	24667	3.74%	
ring_duration	6464	0.98%	Appears only when calls reach the ringing stage (i.e. excludes blocked, failed calls)

**Variables left blank are those with no inherent explanation behind NA values; may be truly missing*

All Calls Data - Categorical Variables

Variable	Total Categories	Example Values
start_time	483703	2024-04-15T20:33:33.465Z (4); 2024-05-02T21:11:03.572Z (4); … [Total: 483703 categories]
answer_time	375804	2025-01-29T03:19:19.448Z (8); 2025-01-29T13:17:56.688Z (8); … [Total: 375804 categories]
called_number	32048	13123478300 (145455); 13123411070 (138386); … [Total: 32048 categories]
correlation_id	320041	a5552eb3-5806-4cb3-8685-141760d35337 (108); 12aaa6ea-dafe-41c1-96c1-dbd4917dea67 (60); … [Total: 320041 categories]
location	1	HQ (660413)
inbound_trunk	10	wcc_Pc_tp-ipRwm_ku064NHZiw (71955); wcc_aoKpA8bhRyueUuXaEF1grQ (1081); … [Total: 10 categories]
outbound_trunk	9	wcc_Pc_tp-ipRwm_ku064NHZiw (138246); wcc_lyq3fhu9TjS8hbku8c4Zcg (5786); … [Total: 9 categories]
direction	2	TERMINATING (384227); ORIGINATING (276186)
call_type	11	SIP_ENTERPRISE (341440); SIP_INBOUND (213566); … [Total: 11 categories]
client_type	4	SIP (331641); WXCC (172700); TEAMS_WXC_CLIENT (95527); WXC_DEVICE (60545)

All Calls Data - Categorical Variables, cont.

Variable	Total Categories	Example Values
client_version	87	12.0.4 (21631); 12.0.5 (13367); … [Total: 87 categories]
sub_client_type	1	MOBILE_APP (43631)
os_type	3	windows (51822); ios (40367); android (3264)
device_mac	262	D4AD71BF7E00 (8826); 002F5C7AC10C (7801); … [Total: 262 categories]
model	8	7841 (41807); 8841 (17667); 7841-3PCC (303); 8841-3PCC (136); … [Total: 8 categories]
answered	2	TRUE (629994); FALSE (30419)
international_country	6	MX (88); GT (3); AU (1); … [Total: 6 categories]
original_reason	6	FollowMe (138762); NoAnswer (66481); … [Total: 6 categories]
related_reason	8	CallForwardNoAnswer (57688); Deflection (38832); … [Total: 8 categories]
redirect_reason	6	NoAnswer (115371); Deflection (29256); … [Total: 6 categories]
site_main_number	1	13125068649 (660413)
site_timezone	2	-300 (471533); -360 (188880)

All Calls Data - Categorical Variables, cont.

Variable	Total Categories	Example Values
user_type	7	User (241360); WCCAdapter (163984); … [Total: 7 categories]
call_id	223581	BW000241106300125122800561 2@10.71.100.200 (4); BW001247602310125171537572 7@10.71.100.200 (4); … [Total: 350183 categories]
local_session_id	292395	0 (15174); 00000000000000000000000000000000 (11927); 458ce6d300105000a000701f5386d893 (15); … [Total: 459914 categories]
remote_session_id	129194	0 (48263); 00000000000000000000000000000000 (39747); f30a115700105000a000002f5c7ac10c (8); … [Total: 204173 categories]
user_uuid	484	fc1b25ce-47c2-4953-8b36-745c16d01b1b (80643); 3c31d35a-0109-45f3-9712-a552436838fb (37078); … [Total: 484 categories]
org_uuid	1	39f56182-133b-466a-b05d-cb8a43b1c76a (660413)
report_id	652159	00010be0-e83b-3c60-8ed1-fb807f55edb0 (2); 00122812-5a63-340d-a38f-ff6e526814b1 (2); … [Total: 652159 categories]
site_uuid	1	38176ce1-16a1-4ab7-b156-5747e56b6cea (660413)
releasing_party	3	Remote (512155); Local (148087); Unknown (171)
redirecting_number	369	13123411070 (144184); 13124235938 (36353); … [Total: 369 categories]

All Calls Data - Categorical Variables, cont.

Variable	Total Categories	Example Values
transfer_related_call_id	39348	57348982466:0 (2); 57348982466:0A (2); … [Total: 39348 categories]
call_transfer_time	69396	2025-01-29T13:17:38.388Z (4); 2025-01-29T14:07:31.709Z (4); … [Total: 69396 categories]
user_number	415	13123411070 (210226); 13123478300 (73917); … [Total: 415 categories]
local_call_id	652159	57292359946:0 (2); 57292633076:0 (2); … [Total: 652159 categories]
remote_call_id	336939	57292662596:0A (2); 57292662608:0 (2); … [Total: 336939 categories]
network_call_id	303655	BW000006591300125500598518@10.71.100.200 (2); BW000819454290125-1276525896@10.71.100.200 (2); … [Total: 303655 categories]
related_call_id	82890	38480982488:0 (21); 63020951178:0 (20); … [Total: 82890 categories]
call_outcome	3	Success (651820); Refusal (7814); Failure (779)
call_outcome_reason	15	Normal (63592); CallRejected (4700); UnassignedNumber (1675); … [Total: 15 categories]

All Calls Data - Categorical Variables, cont.

Variable	Total Categories	Example Values
final_local_session_id	445809	0 (27095); 00000000000000000000000000000000 (22507); 458ce6d300105000a000701f5386d893 (15); … [Total: 445809 categories]
final_remote_session_id	177258	0 (63833); 00000000000000000000000000000000 (53252); 7ee14839e2d44fd7b000ca7b8df885c7 (10); … [Total: 177258 categories]
answer_indicator	3	Yes (548220); Yes-PostRedirection (81774); No (30419)
release_time	387120	2025-01-29T13:18:29.648Z (12); 2025-01-29T14:08:57.415Z (12); 2025-01-29T14:30:35.515Z (12); 2025-01-29T14:32:07.818Z (12); 2025-01-29T15:07:30.264Z (12); … [Total: 387120 categories]
report_time	387120	2025-01-29T13:18:29.648Z (12); 2025-01-29T14:08:57.415Z (12); 2025-01-29T14:30:35.515Z (12); 2025-01-29T14:32:07.818Z (12); 2025-01-29T15:07:30.264Z (12); … [Total: 387120 categories]
pstn_vendor_name	1	CallTower (294655)
pstn_provider_id	1	afc59c71-23c9-4884-bab9-535f916eb11b (300837)
month	11	February 2025 (63669); August 2024 (63262); May 2024 (62944); … [Total: 7 categories]
external_caller_id_number	336	13123411070 (98986); 13123478300 (19010); 13124235938 (9391); … [Total: 336 categories]

All Calls Data - Categorical Variables, cont.

Variable	Total Categories	Example Values
device_owner_uuid	39	7bc5395b-012a-41fc-a044-8c861ce9e40d (1499); 203ac550-21ab-4080-872b-4ff92f816fad (1295); 8eff0dfd-d0c1-4e16-9bdc-5c3d93ae6c28 (431); … [Total: 39 categories]
redirecting_party_uuid	292	3c31d35a-0109-45f3-9712-a552436838fb (10446); d389c857-73fd-4a1f-9179-0392b66abd50 (4578); 0624538c-12e1-44c5-917f-7cf8f45973e3 (3958); … [Total: 292 categories]
user	313	Voicemail (6796); legalclinics VM (2931); LAC Directory English (2253); Front Desk Operator (1897); Bankruptcy Helpdesk (1188); … [Total: 313 categories]