*Analyst: Michelle Tirado*

*January 1st, 2026*

# Final Project– Data Analyst

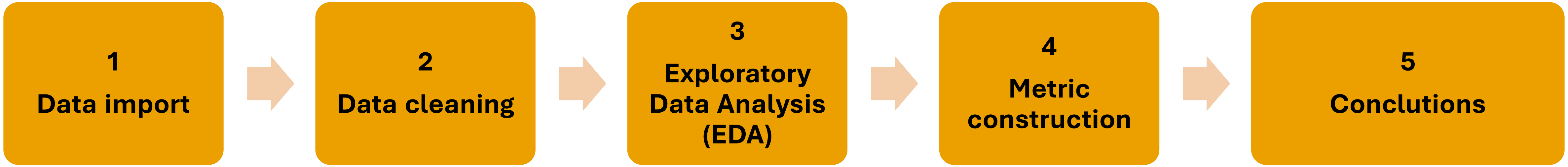
# Project Information

**Date:** January 2nd, 2026.
**Analyst name:** Michelle Tirado
**Proyect name:** “CallMeMaybe”
**Project objective:** Analyze agent performance metrics to identify top-performing and underperforming agents.

# RESUME

This business report presents an analysis of agent performance in a call center environment. The tables, figures, analyses, and conclusions are derived from the data analysis. The data were imported from a CSV file and cleaned to remove duplicate records and empty cells. The cleaned dataset was then analyzed to examine client behavior, operator performance, and to calculate key performance indicators (KPI).

# PIPELINE FOR DATA ANALYSIS



1
Data import

2
Data cleaning

3
Exploratory Data Analysis (EDA)

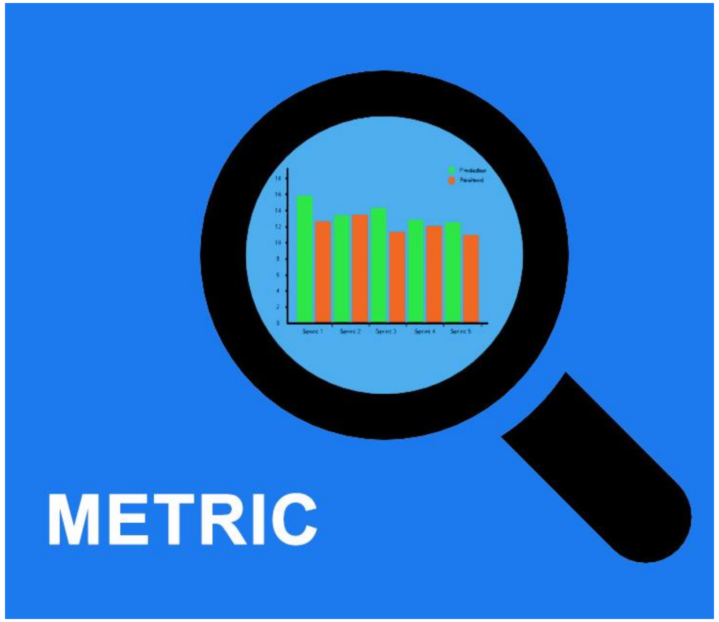
4
Metric construction

5
Conclutions

# DATA DESCRIPTION

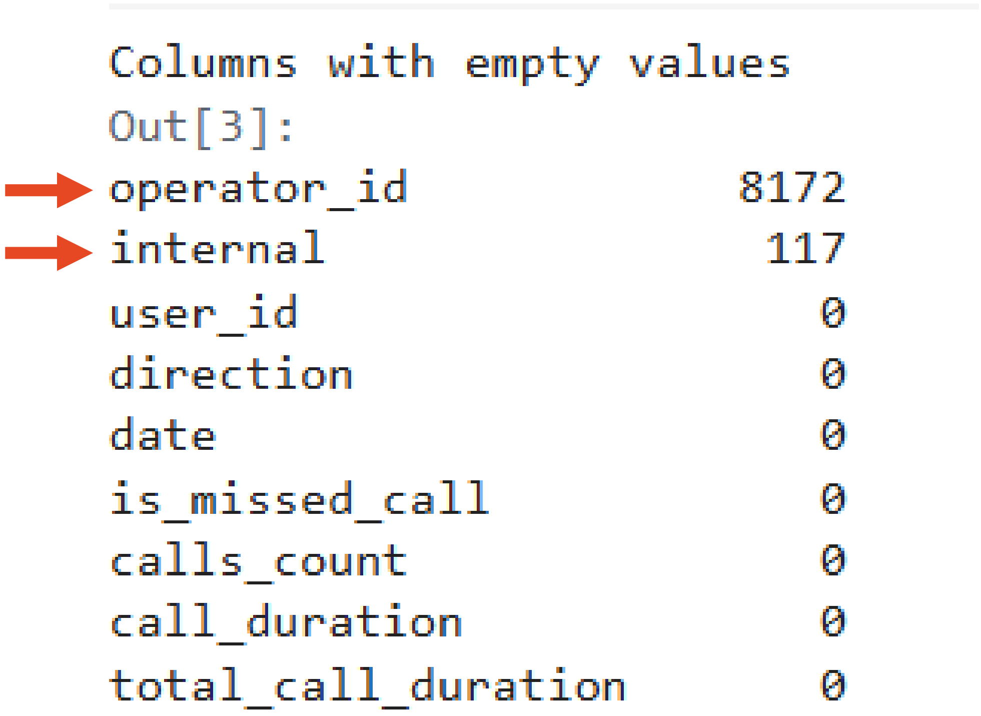
The “telecom_dataset_us.csv” dataset contains information about the usage of the “Call Me Maybe” virtual telephony service. The customers are organizations that need to distribute large volumes of inbound calls among multiple operators. Operators can also place internal calls to communicate with one another. All calls are routed through the Call Me Maybe network. The compressed dataset telecom_dataset_us.csv includes the following columns:

- **user_id:** Unique identifier of the customer (client) account.
- **date:** Date on which the statistics were recorded.
- **direction:** Call direction, either inbound (in) or outbound (out).
- internal: Indicates whether the call was internal (between operators) or external (with a customer).
- **operator_id:** Unique identifier of the operator handling the call.
- **is_missed_call:** Indicates whether the call was missed (True) or successfully answered (False).
- **calls_count:** Number of calls recorded on the given date.
- **call_duration:** Duration of the call in seconds, excluding waiting time.
- **total_call_duration:** Total call duration in seconds, including waiting time.



# 1. DATA IMPORT

The dataset contains 8 columns and a total of 53,902 entries. The "operator_id" and "internal" columns contain missing values, which require further investigation during the data-cleaning process. In addition, some columns have incorrect data types that need to be corrected as part of data cleaning.



```
Columns with empty values
Out[3]:
operator_id            8172
internal                117
user_id                   0
direction                 0
date                      0
is_missed_call            0
calls_count               0
call_duration             0
total_call_duration       0
```

# 2. DATA CLEANING

This phase includes: A) Cell format correction, B) Removing empty cells, and C) Eliminating duplicates.

**The original data has:**

Total entries: 53,902

Total users: 307

Total operators: 1,092

Total number of recorded dates: 119

**Missing information:**

Total empty entries: 8,172

Total number of affected users: 305

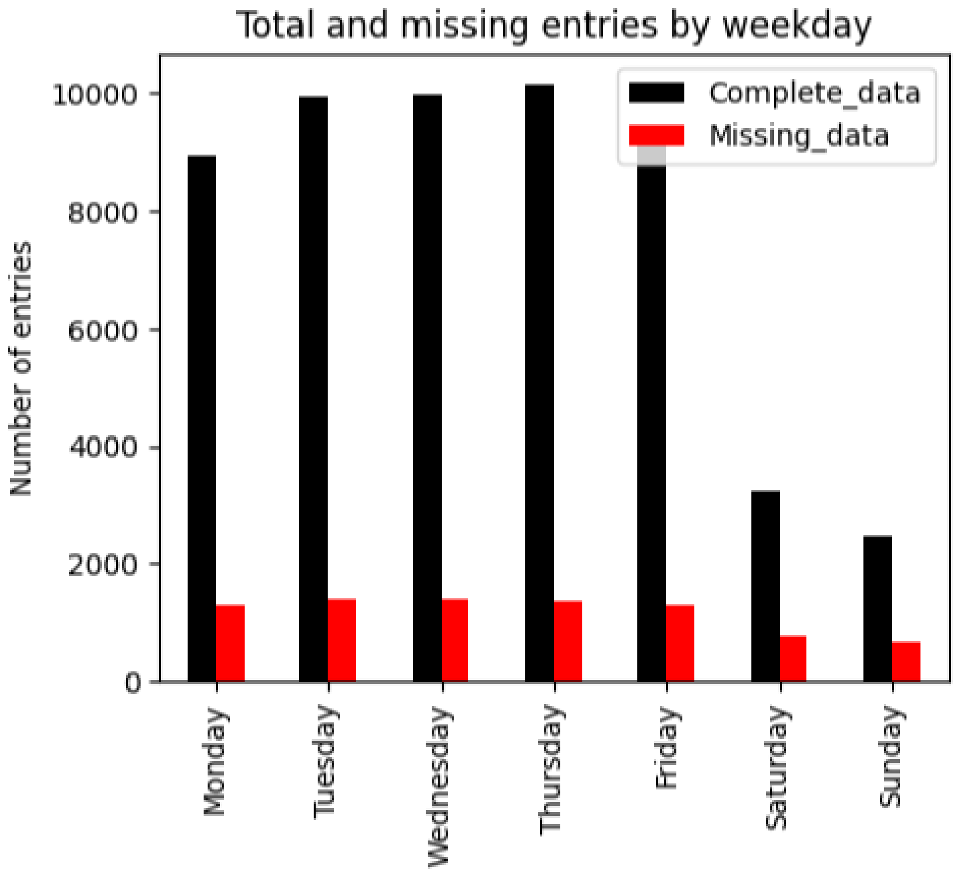
Total number of affected days: 119

**Duplicates:**

Duplicated cells: 4,179

Empty cells were found in the **operator_id** field, which is a categorical variable. Since these values cannot be re placed, the corresponding records were removed.

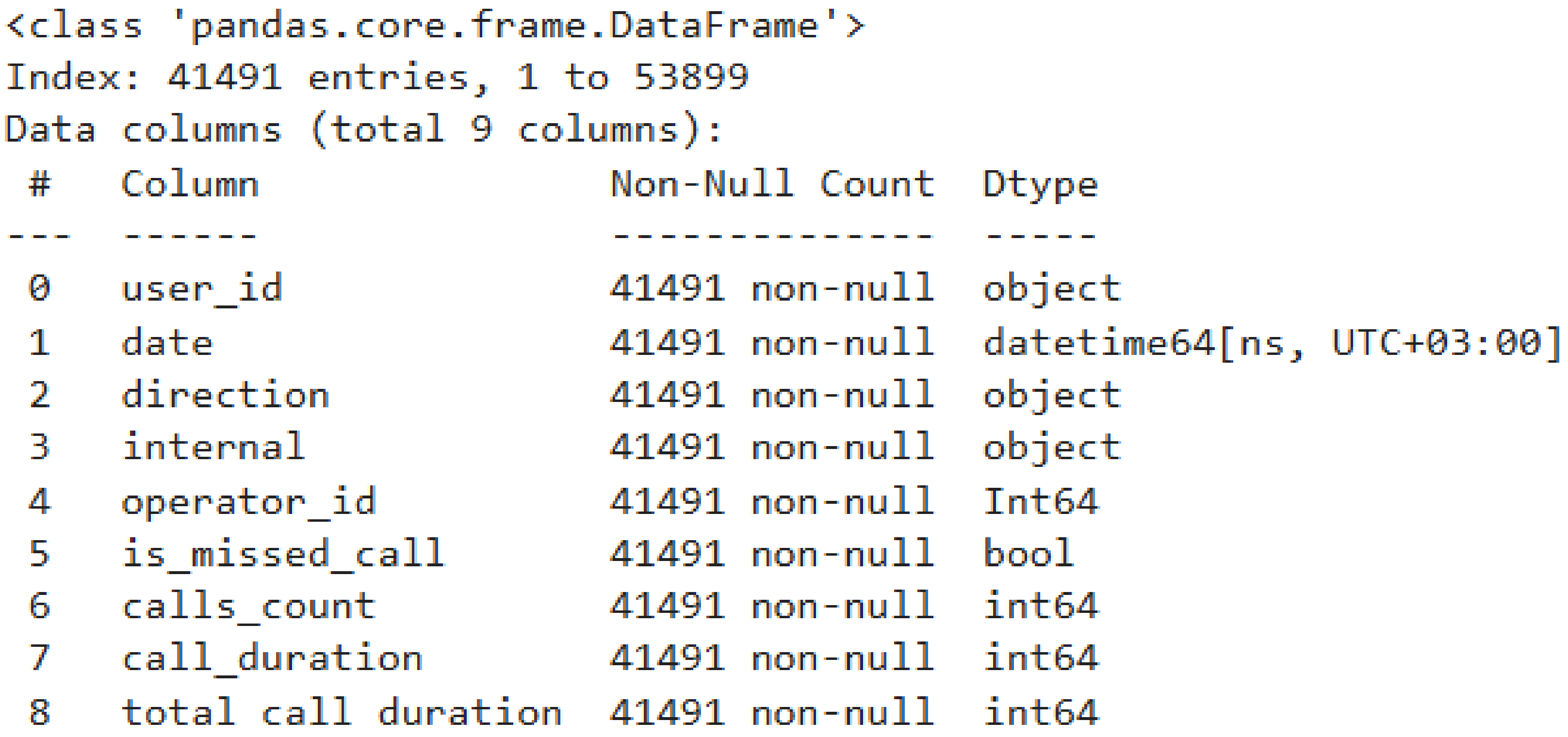
**The missing entries represent 15.16% of the total entries.** To identify the most impacted days, a graph was constructed. Call volume is higher on weekdays compared to weekends; consequently, the amount of missing data is greater on these days. However, when calculating the proportion of missing cells recorded per day (named missing rate), it becomes evident that **weekends have the highest missing data rates.**



| | Missing rate (%) |
|---|---|
| Monday | 14.26 |
| Tuesday | 14.13 |
| Wednesday | 14.03 |
| Thursday | 13.48 |
| Friday | 14.01 |
| Saturday | 23.53 |
| Sunday | 27.40 |

# CLEAN DATASET

The cleaned dataset was reduced to **41,491** entries.



```
<class 'pandas.core.frame.DataFrame'>
Index: 41491 entries, 1 to 53899
Data columns (total 9 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   user_id              41491 non-null  object
 1   date                 41491 non-null  datetime64[ns, UTC+03:00]
 2   direction            41491 non-null  object
 3   internal             41491 non-null  object
 4   operator_id          41491 non-null  Int64
 5   is_missed_call       41491 non-null  bool
 6   calls_count          41491 non-null  int64
 7   call_duration        41491 non-null  int64
 8   total call duration  41491 non-null  int64
```

# 3. EXPLORATORY DATA ANALYSIS (EDA)

Basic statistic was used to explore data, and the greatest finding are presented:

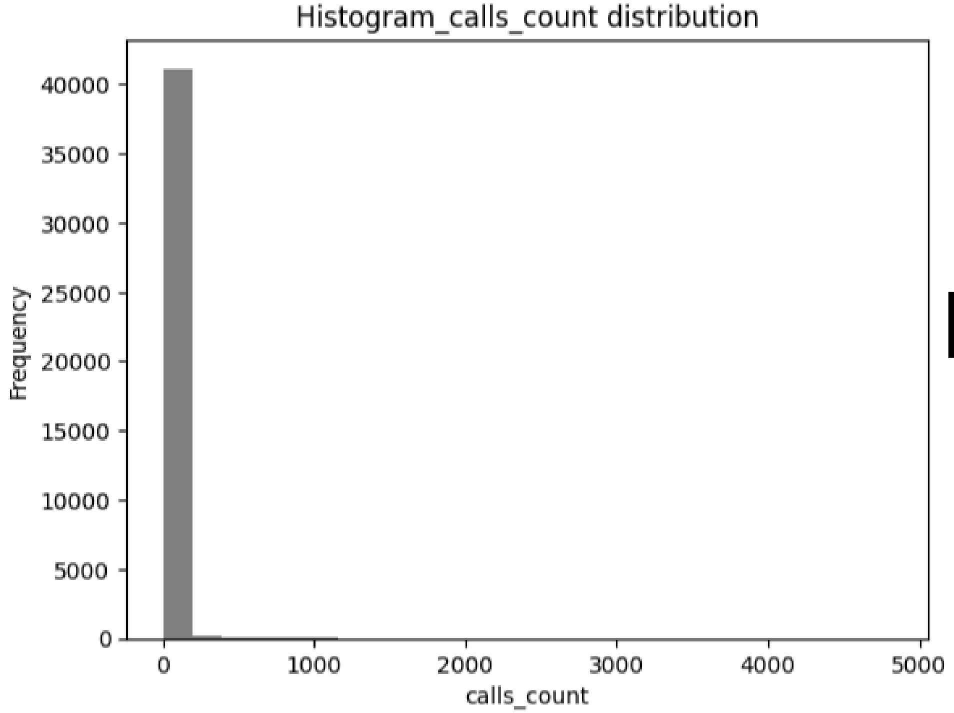
- The minimum date recorded in the dataset is 2019-08-02
- The maximum date recorded in the dataset is 2019-11-28
- The date range recorded in the dataset is 118 days

- The minimum calls count recorded by agent is 1
- The maximum calls count recorded by agent is 4,817
- The average calls count is 17
- The median calls count is 4.0

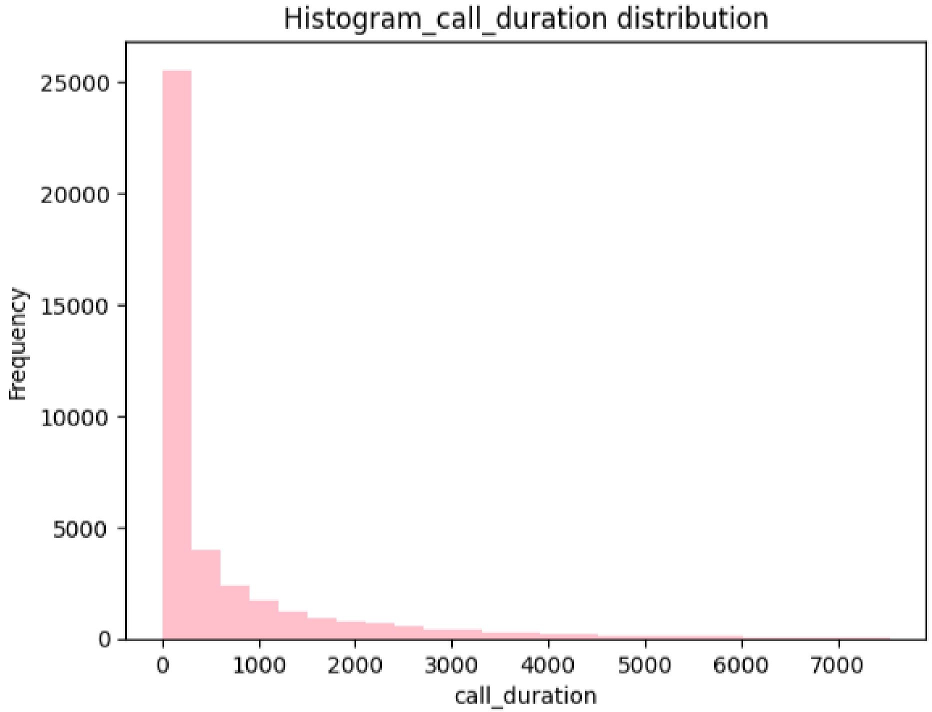
- The minimum call duration recorded in the dataset is 0 seconds
- The maximum call duration recorded in the dataset is 144,395 seconds (40 hours)
- The call duration mean recorded in the dataset is 1,011 seconds (17 minutes)
- The call duration median recorded in the dataset is 106 (~ 2 minutes)

- The minimum waiting duration recorded in the dataset is 0 seconds
- The maximum waiting duration recorded in the dataset is 46,474 (13 hours)
- The waiting_duration mean recorded in the dataset is 312 seconds
- The waiting_duration median recorded in the dataset is 60 seconds

The exploratory analysis reveals a **highly right-skewed distribution**. For example, the maximum waiting duration recorded in the dataset is 46,474 seconds (approximately 13 hours), which makes no sense, and likely indicates a failure in the recording system for that observation. Therefore, a data **trimming process** was applied to remove extreme values in *calls_count*, *call_duration*, and *total_call_duration* as all these variable were affected. As a result, only the 0–99th percentile of the data was retained to reduce the impact of highly right-skewed distributions.



**TRIMMING PROCESS**

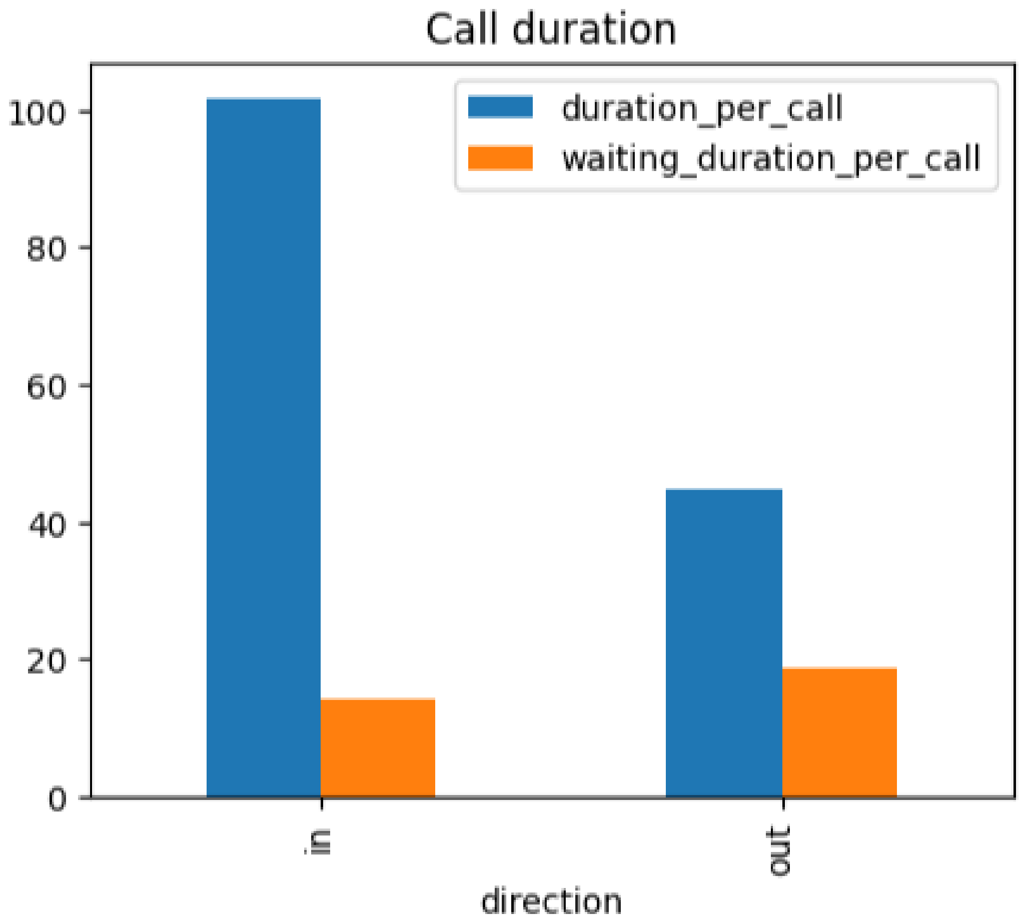


# RESULTS OF THE TRIMMING PROCESS

## INSIGHTS

| DATASET | | TRIMMED DATASET |
|---|---|---|
| The maximum number of calls recorded per agent per day in the non-trimmed dataset is 4,817. | → | The maximum number of calls recorded per agent per day in the trimmed dataset is 164. |
| The longest call recorded in the non-trimmed dataset dataset is 40.11 hours. | | The longest call recorded in the trimmed dataset is 2.09 hours. |
| The longest waiting time recorded in the non-trimmed dataset is 12.91 hours. | | The longest waiting time recorded in the trimmed dataset is 1.56 hours. |

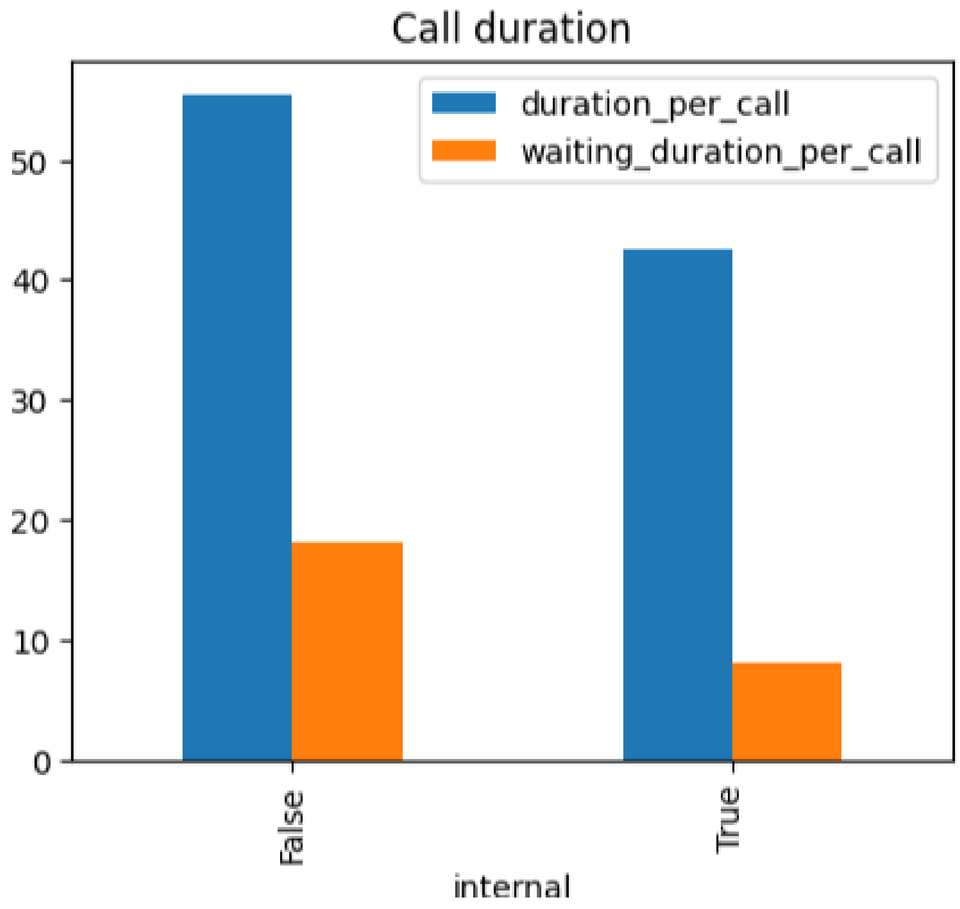
# IINBOUND CALLS VS OUTBOUND CALLS



| | INBOUND | OUTBOUND |
|---|---|---|
| Total calls | 80,535 | 367,885 |
| Duration per call (seconds) | 102 | 45 |
| Waiting time per call (seconds) | 14 | 19 |

In the database, there is a record of 17.96% incoming calls and 82.04% outbound calls.

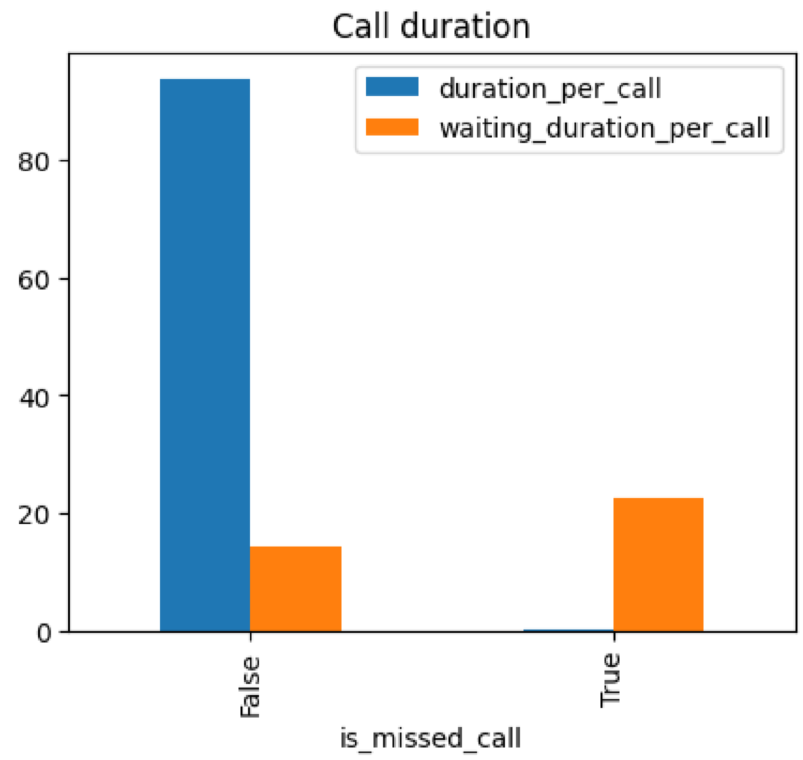
# INTERNAL VS EXTERNAL CALLS



| | EXTERNAL | INTERNAL |
|---|---|---|
| Total calls | 434,755 | 13,665 |
| Duration per call (seconds) | 56 | 18 |
| Waiting time per call (seconds) | 43 | 8 |

There is 96.95% external calls and 3.05% internal calls

# MISSED CALL VS ANSWERED CALL



| | MISSED CALL | ANSWERED CALL |
|---|---|---|
| Total calls | 184,962 | 263,458 |
| Duration per call (seconds) | 0 | 94 |
| Waiting time per call (seconds) | 23 | 14 |

There is 58.75% received calls and 41.25% missed calls.

# Exploratory Data Analysis Insights

- Overall, operators spend more time on outbound calls than on inbound calls. The total call duration for outbound calls is approximately twice that of inbound calls.

- Most calls (96.95%) involve external sources (clients). Of all calls, 58.75% are answered and 41.25% are missed. Missed calls have a higher average waiting time (23 seconds) compared to answered calls (14.36 seconds).

- On average, an operator spends 93.69 seconds per call, with an average waiting time of 14.36 seconds. Only 63.54% of calls last longer than 10 seconds.

# 4. METRIC CONSTRUCTION

Using the trimmed dataset, **Key Performance Indicators (KPIs)** will be constructed. Based on standard teleperformance KPIs found in external bibliographic resources, a dataset containing these indicators will be built. The selected metrics are:

- **Total calls**
- **Average Handle Time (AHT): Average time per call (seconds)**
- **Average Waiting Time (AWT): Average waiting time per call (seconds)**
- **Abandon Rate (AR): Is calculated as missed_calls divided by total_calls, with values ranging from 0 to 1.**

Agent performance is evaluated using AHT, AWT, and AR, where lower values indicate better performance, while higher values reflect poorer performance. Because this metrics are independent variables, they will be standardized to be summed and create a **performance index**. The proposed formula is:

$$Performance\ Index = zAHT + zAWT, + zAR$$

Where:

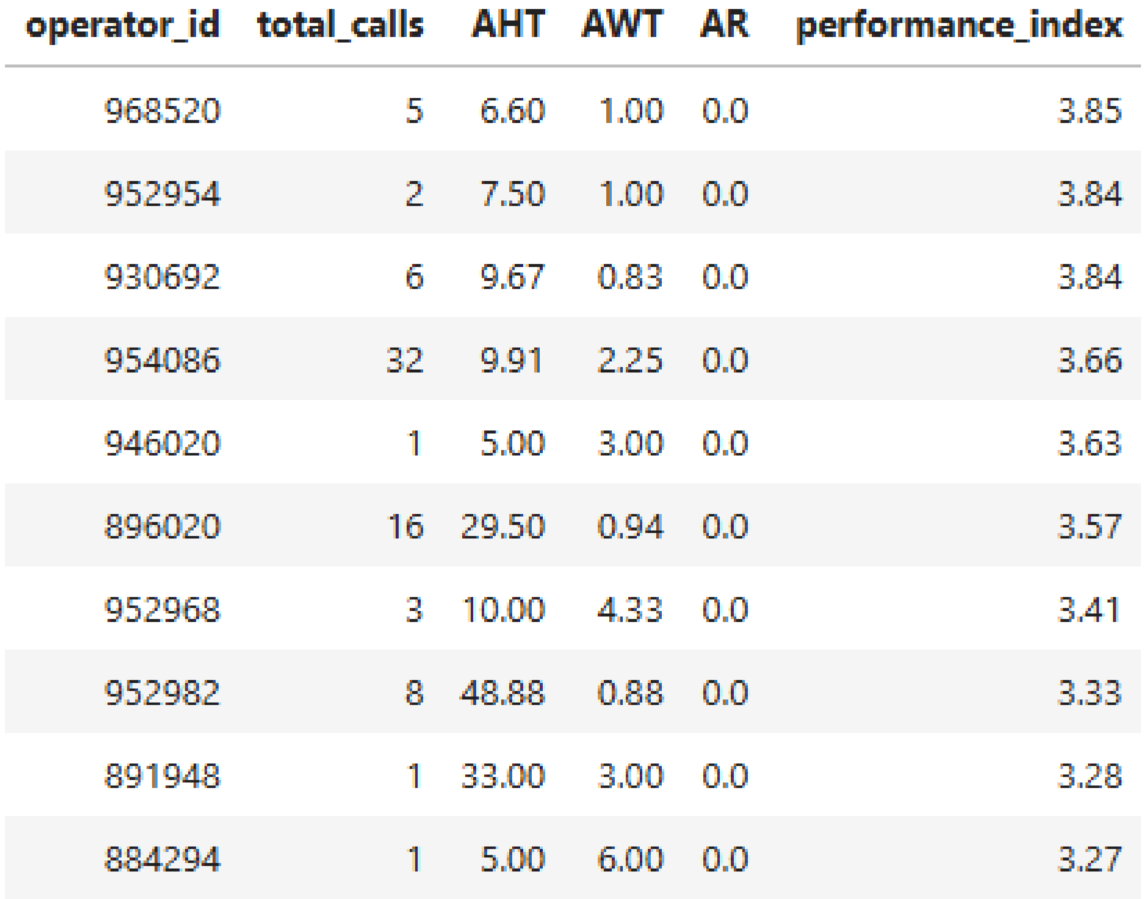
$zAHT$ = standarized Average Handle Time

$zAWT$ = standarized Average Handling Time

$zAR$ = standarized Abandon Rate

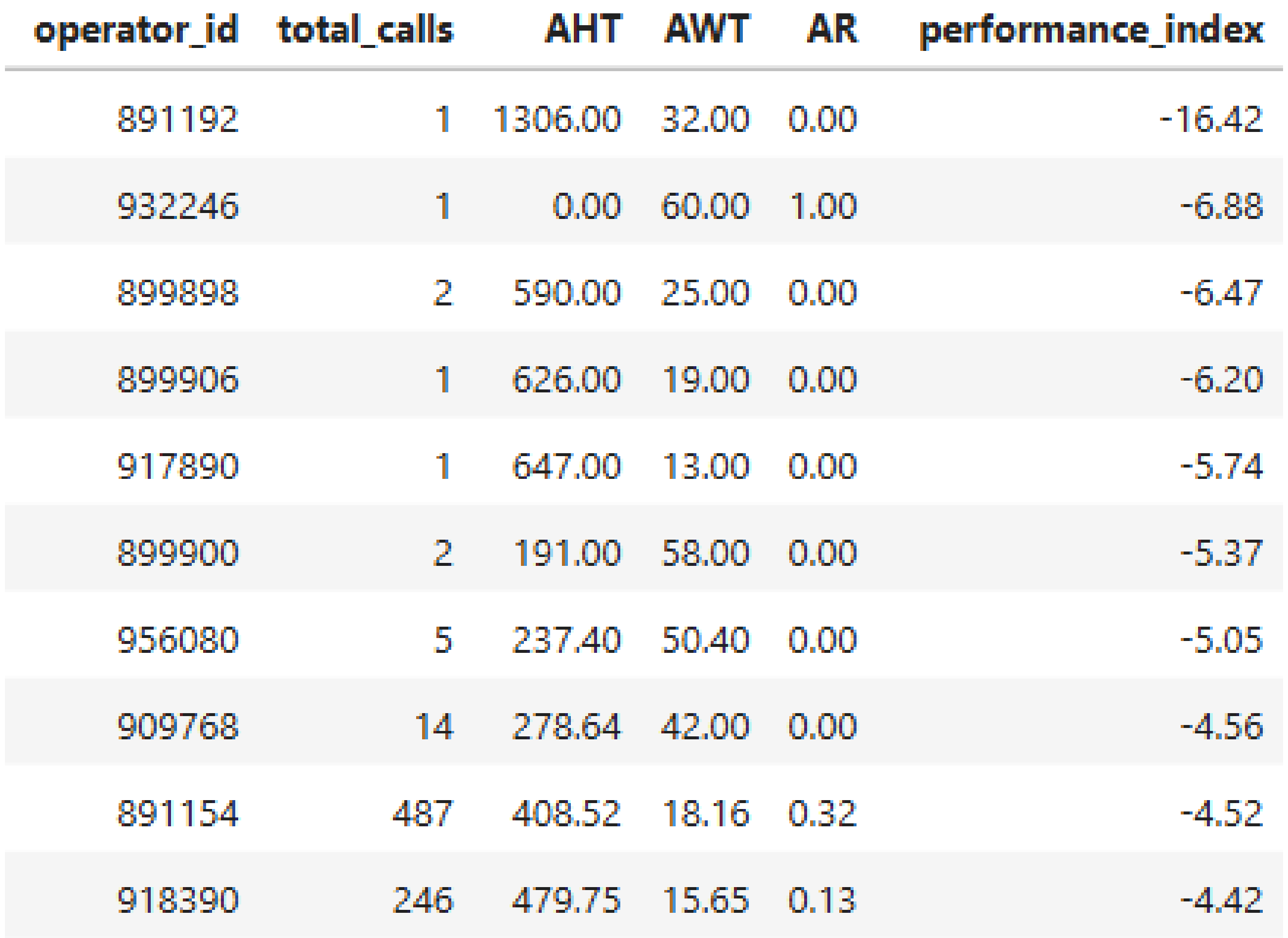
To make this report more effective, only the 10 highest- and lowest-performing operators are presented. The calculated metrics for all operators are included in the “agents_performance.xlsx” file.

# 10 BEST PERFORMANCE OPERATORS



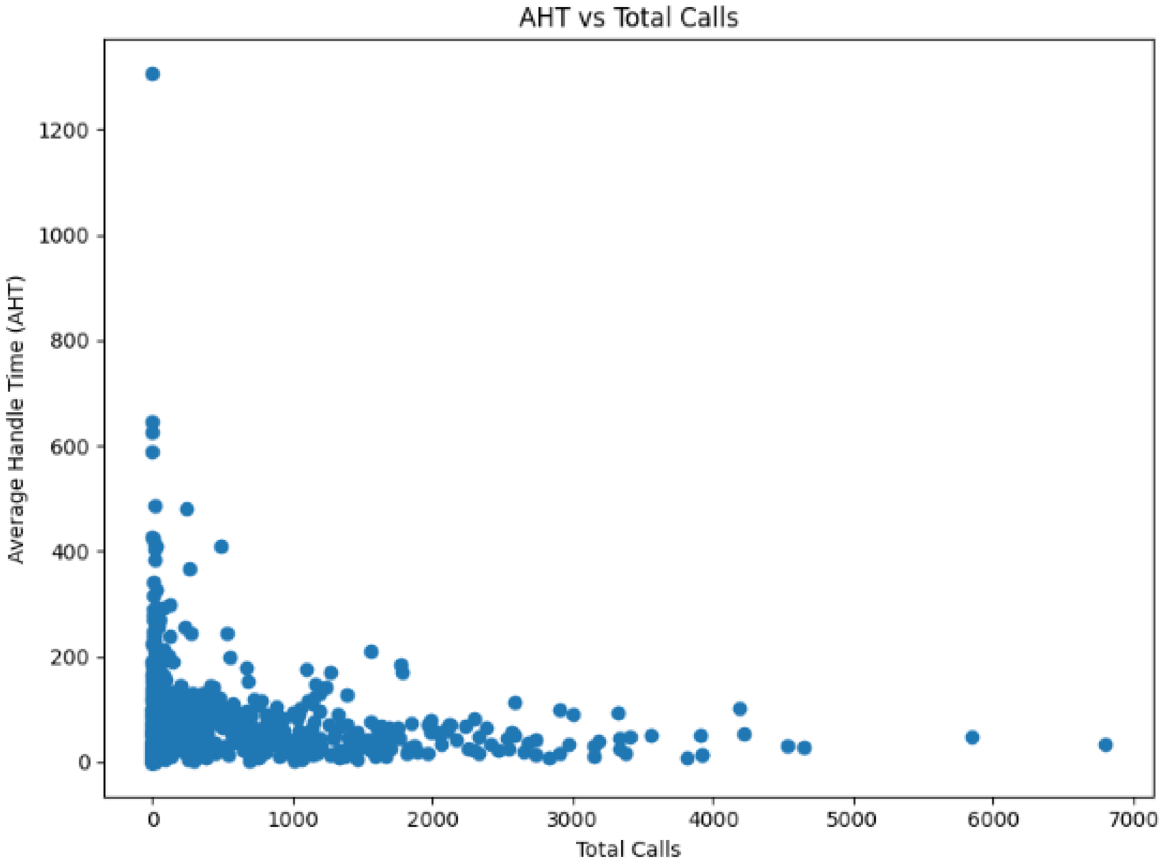
| operator_id | total_calls | AHT | AWT | AR | performance_index |
|---:|---:|---:|---:|---:|---:|
| 968520 | 5 | 6.60 | 1.00 | 0.0 | 3.85 |
| 952954 | 2 | 7.50 | 1.00 | 0.0 | 3.84 |
| 930692 | 6 | 9.67 | 0.83 | 0.0 | 3.84 |
| 954086 | 32 | 9.91 | 2.25 | 0.0 | 3.66 |
| 946020 | 1 | 5.00 | 3.00 | 0.0 | 3.63 |
| 896020 | 16 | 29.50 | 0.94 | 0.0 | 3.57 |
| 952968 | 3 | 10.00 | 4.33 | 0.0 | 3.41 |
| 952982 | 8 | 48.88 | 0.88 | 0.0 | 3.33 |
| 891948 | 1 | 33.00 | 3.00 | 0.0 | 3.28 |
| 884294 | 1 | 5.00 | 6.00 | 0.0 | 3.27 |

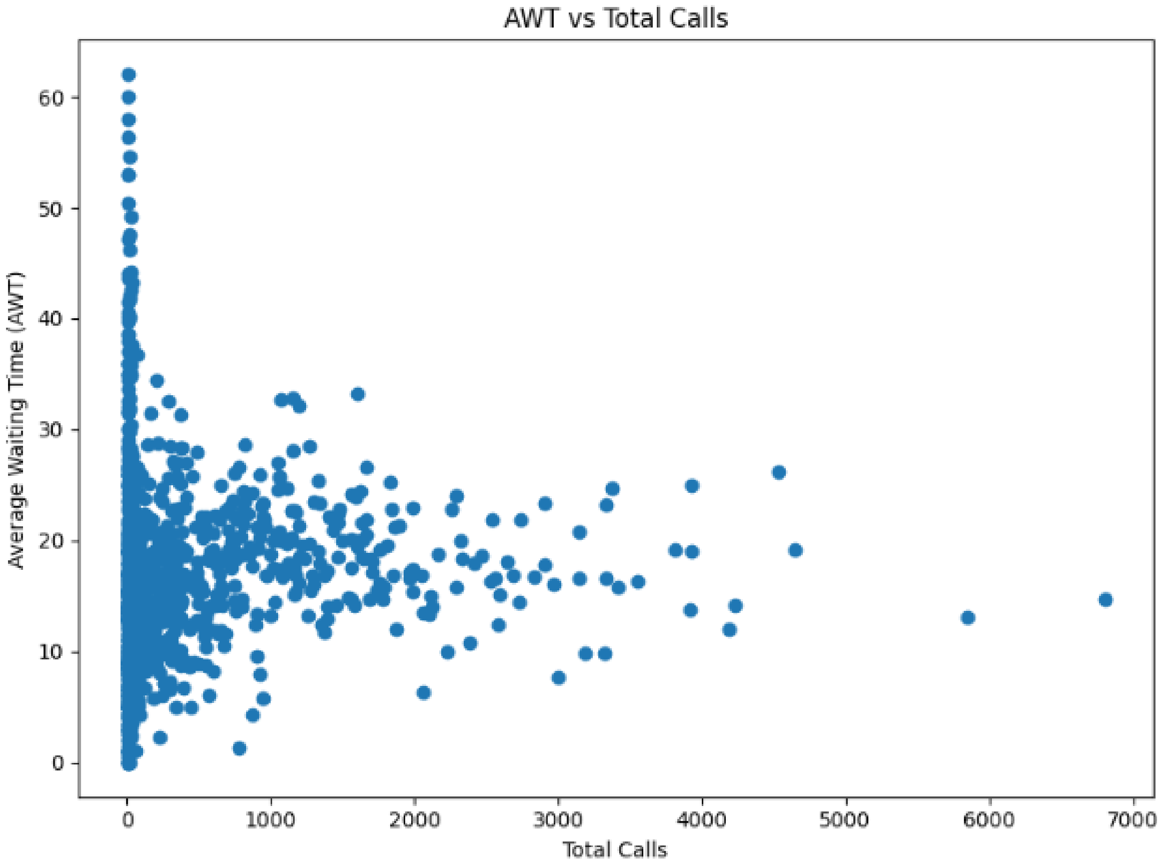
# 10 LOWEST PERFORMANCE OPERATORS



| operator_id | total_calls | AHT | AWT | AR | performance_index |
|---:|---:|---:|---:|---:|---:|
| 891192 | 1 | 1306.00 | 32.00 | 0.00 | -16.42 |
| 932246 | 1 | 0.00 | 60.00 | 1.00 | -6.88 |
| 899898 | 2 | 590.00 | 25.00 | 0.00 | -6.47 |
| 899906 | 1 | 626.00 | 19.00 | 0.00 | -6.20 |
| 917890 | 1 | 647.00 | 13.00 | 0.00 | -5.74 |
| 899900 | 2 | 191.00 | 58.00 | 0.00 | -5.37 |
| 956080 | 5 | 237.40 | 50.40 | 0.00 | -5.05 |
| 909768 | 14 | 278.64 | 42.00 | 0.00 | -4.56 |
| 891154 | 487 | 408.52 | 18.16 | 0.32 | -4.52 |
| 918390 | 246 | 479.75 | 15.65 | 0.13 | -4.42 |

As shown in the AHT graphic, users' calls tend to last between 0 and 200 seconds.
As shown in the AWT graphic, users tend to wait between 0 and 30 seconds before their calls are answered.





# CONCLUTIONS

The original dataset contained 53,902 entries. A total of 8,172 rows with missing values in the "operator_id" and "internal" columns were removed, and an additional 4,179 rows were removed due to duplicated records. The missing entries were analyzed to determine whether the missing data was associated with a specific date or client; however, no specific pattern was identified. Weekends exhibited the highest missing-value rates. Overall, the missing entries represented 15.16% of the total dataset and were removed from the analysis. After the data-cleaning process, the dataset was reduced by 23.03%, resulting in a final dataset of 41,491 entries.

The dataset presents the records obtained from 2019-08-02 to 2019-11-28 (118 days).

The clean dataset was trimmed as inconsistent extreme values were detected.

The operators spend more time on outbound calls than inbound calls.

There is 58.75% received calls and 41.25% missed calls. The missed calls has an average record of 23 seconds on waiting time, which is higher than the 14.36 seconds recorded for non-missed calls.

Only 63.54% of the calls last more than 10 seconds, contributing to the observed call duration distribution (right-skewed).

# REFERENCIAS USADAS PARA EL ANÁLISIS

Holmes, A., Illowsky, B., & Dean, S. (14 de Febrero de 2022). (OpenStax) Recuperado el Diciembre de 2025, de Introducción a la estadística empresarial; El coeficiente de correlación r: https://openstax.org/books/introducci%C3%B3n-estad%C3%ADstica-empresarial/pages/13-1-el-coeficiente-de-correlacion-r

Papageorgiou, G., Grant, S. W., Takkenberg, J. J., & Mokhles, M. M. (2018). Statistical primer: how to deal with missing data in scientific research? *Interactive CardioVascular and Thoracic Surgery, 27*(02), 153–158. doi: https://doi.org/10.1093/icvts/ivy102

*Revista Completa*. (17 de Noviembre de 2025). Obtenido de Coeficiente de correlación: concepto y aplicaciones en estadística: https://revistacompleta.com/coeficiente-de-correlacion-de-pearson/

VCC live. (s.f.). *Call abandonment rate*. Recuperado el 12 de Diciembre de 2025, de https://vcc.live/call-center-kpis/call-abandonment-rate/#:~:text=An%20acceptable%20call%20abandonment%20rate%20is%20typhttps://vcc.live/call-center-kpis/call-abandonment%20-rate/#:~:text=An%20acceptable%20call%20abandonment%20rate%20is%20typically,Competitive%20markets%20*%20Contact%20centers%20with%20SLAscally,Competitive%20markets%20*%20Contact%20centers%20with%20SLAs

Yi, M. (s.f.). *ATTLASIAN*. Recuperado el 12 de Enero de 2025, de A complete guide to box plots: https://www.atlassian.com/data/charts/box-plot-complete-guide