Code Logic & Processing Documentation

# Phase 1: Data Foundation

This phase prepares and enriches raw call transcript data.

## Steps & Logic

* Load & Inspect Data – Reads call transcript data from Excel, prints row and unique counts.
* Data Quality Checks – Counts missing transcripts and speaker info, checks duplicates.
* Data Cleaning & Standardization – Strips whitespace, removes empty transcripts, standardizes `speaker`.
* Derived Columns – Calculates word\_count, char\_count, duration\_seconds, relative\_position.
* Sentiment Analysis – Uses TextBlob to classify each line as Positive, Neutral, or Negative.
* Call-Level Aggregation – Summarizes lines, words, duration, talk time % per call.
* Agent-Level Aggregation – Summarizes productivity and sentiment per agent.
* Validation – Checks for missing agents/customers, timing errors, unrealistic durations.
* Visualization – Histograms for duration, lines per call, talk time %, sentiment distribution.
* Export – Saves cleaned dataset and summary CSVs.

# Phase 2: Speaker Timeline Analysis

This phase analyzes conversation flow, interruptions, and patterns using Phase 1 output.

## Steps & Logic

* Load Phase 1 Outputs – Reads CSVs generated from Phase 1.
* Speaker Turn Analysis – Counts turn changes per call and computes turn balance.
* Speaking Time Patterns – Calculates talk time %, words per turn, and balance.
* Interruption Detection – Flags potential interruptions based on duration, words, punctuation, and keywords.
* Timeline Visualization – Plots per-call speaking segments over time with color-coded bars.
* Conversation Flow Heatmap – Generates a heatmap of active speakers across all calls.
* Clustering Patterns – Uses KMeans clustering to identify Balanced, Agent-Dominated, or Customer-Dominated patterns.
* Summary Visualizations – Histograms and pie charts of talk time, balance, interruptions, and patterns.
* Export – Saves detailed analysis and interruption data to CSV.

# Key Calculations

|  |  |  |
| --- | --- | --- |
| Metric | Calculation | Phase |
| word\_count | len(transcript.split()) | Phase 1 |
| duration\_seconds | (endoffset - startoffset)/1000 | Phase 1 |
| agent\_talk\_time\_percent | agent\_duration\_seconds / total\_duration \* 100 | Phase 1 |
| conversation\_balance | abs(agent\_time - customer\_time)/total\_time | Phase 2 |
| interruptions | Count of lines matching interruption conditions | Phase 2 |
| conversation\_pattern | KMeans clustering (3 clusters) | Phase 2 |

# Output Files

|  |  |
| --- | --- |
| File | Description |
| phase1\_cleaned\_data.csv | Line-level cleaned data with metrics and sentiment. |
| phase1\_call\_summary.csv | Call-level metrics (lines, words, talk times). |
| phase1\_agent\_summary.csv | Agent-level productivity & sentiment stats. |
| phase2\_speaker\_analysis.csv | Call-level enriched metrics incl. interruptions & patterns. |
| phase2\_interruptions.csv | Line-level interruption flags with context. |
| phase2\_timeline\_data.csv | Encoded sequences for timeline visualizations. |