Untitled

2025-03-28

```
df <- read.csv('/Users/joyceli/Desktop/Training.csv')

df <- df %>%
  rename(match_id = "Match.ID.18Char",
         completion_date = "Completion.Date",
         closure_date = "Match.Closure.Meeting.Date",
         contact_notes = "cleaned_notes")

df <- df %>%
  mutate(
    Completion.Date = as.Date(completion_date),
    Match.Closure.Meeting.Date = as.Date(closure_date),

    # Calculate months_to_closure - time from log to closure date in months
    # For rows where Stage == 0 (closed matches), calculate the difference
    # For rows where Stage == 1 (active matches), set to NA
    months_to_closure = case_when(
      Stage == 1 & !is.na(completion_date) & !is.na(closure_date) ~
        round(as.numeric(interval(completion_date, closure_date) / months(1)), 1),
      Stage == 0 ~ NA_real_,  # Set to NA for active matches
      TRUE ~ NA_real_  # Handle any other cases (missing dates, etc.)
    )
  )

df <- df %>%
  mutate(closing_soon = ifelse(months_to_closure <= 3 & months_to_closure > 0, 1, 0))

df <- df %>%
  mutate(closing_soon = case_when(
    # Match will end within 3 months
    months_to_closure <= 3 & months_to_closure > 0 ~ 1,
    # Match will continue longer than 3 months
    months_to_closure > 3 ~ 0,
    # No closure date information or already closed
    TRUE ~ 0
  ))
# 1 is late stage
# 0 is early stage
```

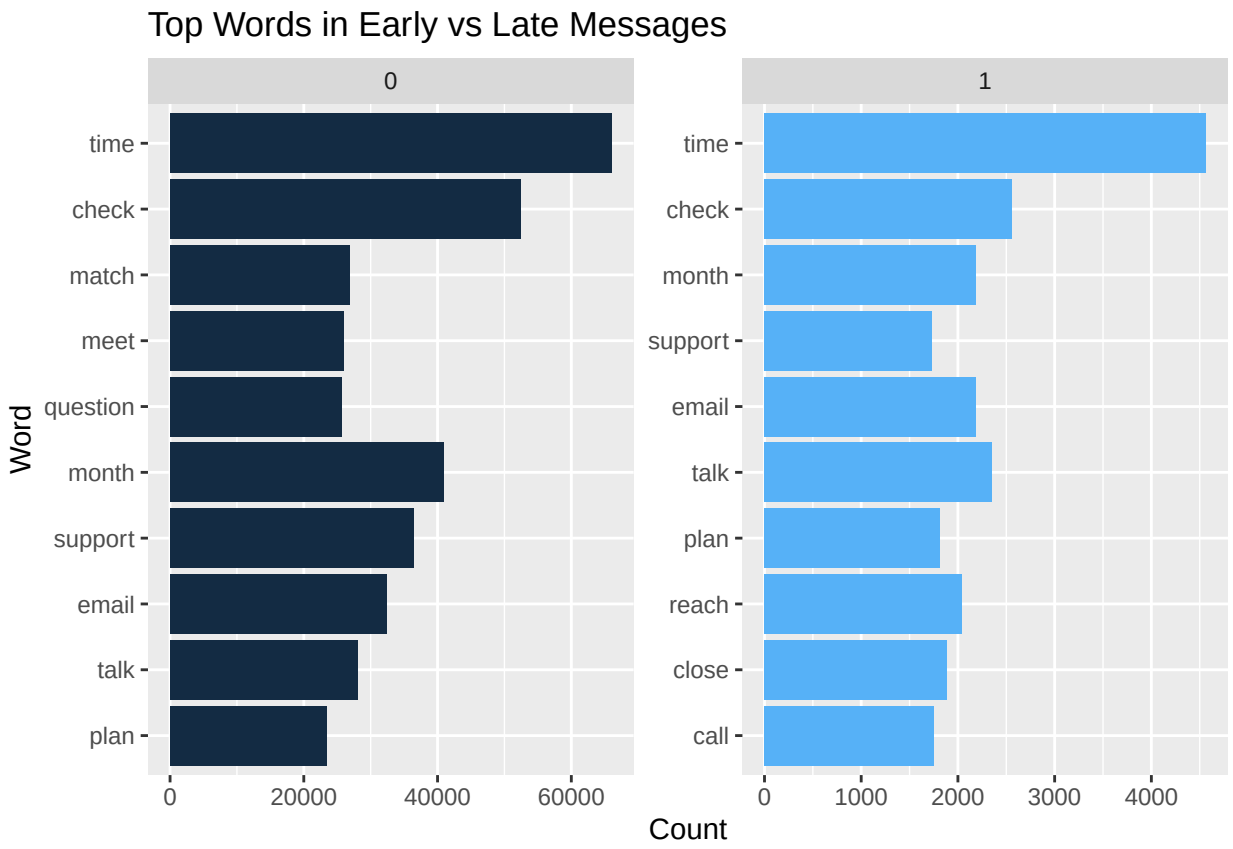
Examine word choice, topics, and tone in early vs late communications

```
# Tokenize text and remove stop words
word_counts <- df %>%
  unnest_tokens(word, contact_notes) %>%
  anti_join(stop_words) %>%
  count(closing_soon, word, sort = TRUE)
```

```
## Joining with `by = join_by(word)`
```

```
# Visualize word differences
library(ggplot2)

word_counts %>%
  group_by(closing_soon) %>%
  top_n(10, n) %>%
  ggplot(aes(x=reorder(word, n), y=n, fill=closing_soon)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~closing_soon, scales = "free") +
  coord_flip() +
  labs(title="Top Words in Early vs Late Messages", x="Word", y="Count")
```




```
# sentiment analysis
joy_words <- c("happy", "enjoy", "joy", "fun", "love", "smile", "laugh", "excite",
               "great", "well", "good", "nice", "awesome", "wonderful", "fantastic")

trust_words <- c("trust", "honest", "loyal", "respect", "reliable", "depend",
                 "faith", "believe", "confident", "committed", "responsible")

fear_words <- c("fear", "afraid", "worry", "scared", "anxious", "nervous",
                "concern", "stress", "danger", "risk", "threat")

anger_words <- c("anger", "angry", "mad", "hate", "rage", "irritate", "annoyed",
                 "frustrate", "upset", "bitter", "hostile", "aggressive")
```

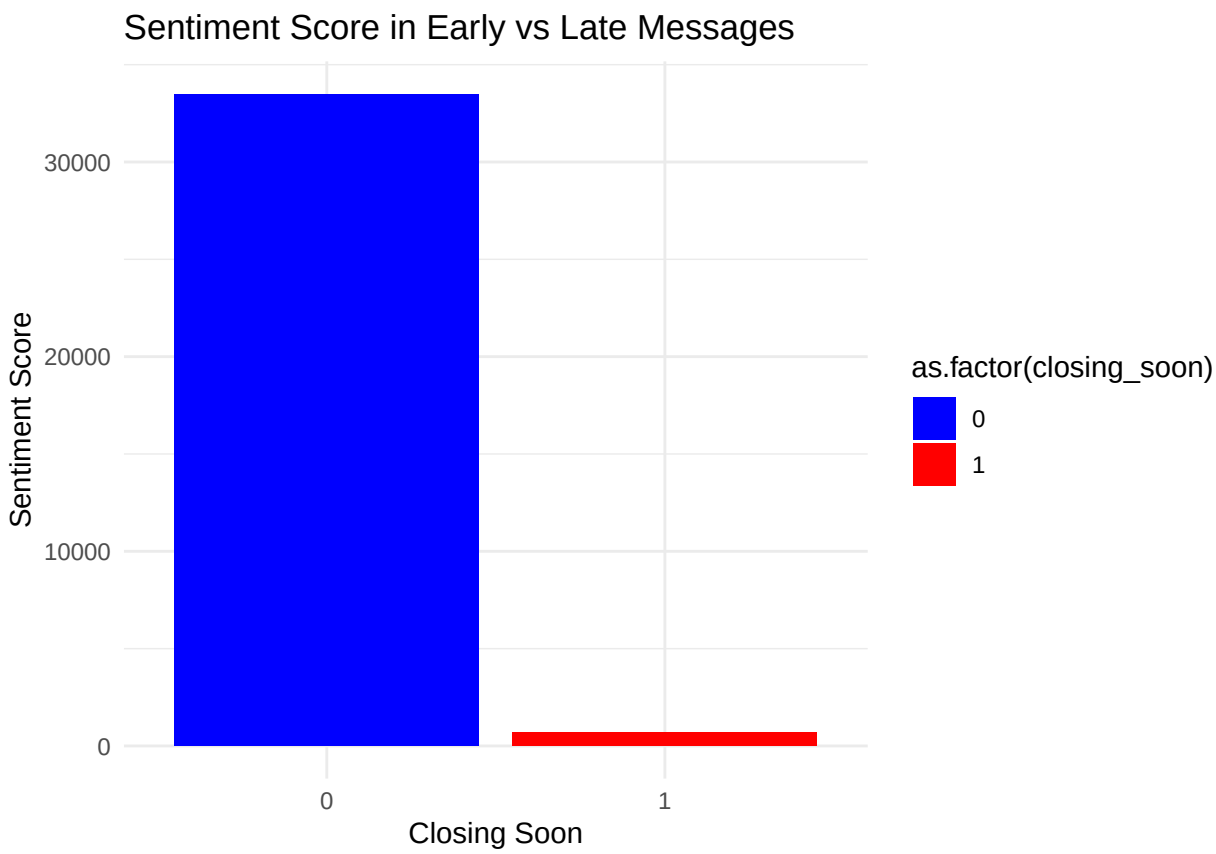
```r
sadness_words <- c("sad", "unhappy", "depress", "sorry", "grief", "disappoint",
                   "miss", "hurt", "pain", "cry", "tear", "alone", "lonely")

# Positive emotions
positive_words <- c(joy_words, trust_words)

# Negative emotions
negative_words <- c(fear_words, anger_words, sadness_words)

# Perform sentiment analysis
sentiment_scores <- df %>%
  unnest_tokens(word, contact_notes) %>%
  mutate(sentiment = case_when(
    word %in% positive_words ~ "positive",
    word %in% negative_words ~ "negative",
    TRUE ~ "neutral"
  )) %>%
  count(closing_soon, sentiment) %>%
  spread(sentiment, n, fill = 0) %>%
  mutate(sentiment_score = positive - negative)

# Visualize sentiment scores for early vs late messages
ggplot(sentiment_scores, aes(x=as.factor(closing_soon), y=sentiment_score, fill=as.factor(closing_soon)))
  geom_bar(stat="identity") +
  labs(title="Sentiment Score in Early vs Late Messages", x="Closing Soon", y="Sentiment Score") +
  scale_fill_manual(values=c("blue", "red")) +
  theme_minimal()
```



Identify language patterns that predict match success or failure

Comparing Language Evolution in Successful vs. Unsuccessful Matches

Successful match?

```
# A match is successful if:
# 1. It's active
# 2. It has a long duration
# 3. Closure reason is "Success"

df$successful_match <- FALSE  # Initialize as unsuccessful

# Mark as successful for different conditions
df$successful_match[df$Stage == 0] <- TRUE
df$successful_match[df$Closure_Reason_Category == "Success"] <- TRUE

# Mark long duration matches as successful
long_duration_threshold <- quantile(df$Match.Length, 0.8, na.rm = TRUE)
df$successful_match[df$Match.Length > long_duration_threshold & !is.na(df$Match.Length)] <- TRUE

# Calculate the success rate
success_rate <- mean(df$successful_match, na.rm = TRUE)
print(paste("Overall success rate:", round(success_rate * 100, 1), "%"))
```

```
## [1] "Overall success rate: 41.5 %"
```

```r
# Categorize match status
df <- df %>%
  mutate(match_status = case_when(
    !is.na(closure_date) ~ "successful",
    is.na(closure_date) ~ "unsuccessful",
    TRUE ~ "unknown"
  ))
```

```r
# Tokenize text and compare word usage across time
word_evolution <- df %>%
  unnest_tokens(word, contact_notes) %>%
  anti_join(stop_words) %>%
  count(match_status, word, sort = TRUE)
```

```
## Joining with `by = join_by(word)`
```

```r
# Visualize word usage differences in successful vs unsuccessful matches
word_evolution %>%
  group_by(match_status) %>%
  top_n(10, n) %>%
  ungroup() %>%
  ggplot(aes(x=reorder(word, n), y=n, fill=match_status)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~match_status, scales = "free") +
  coord_flip() +
  labs(title="Top Words in Successful vs Unsuccessful Matches", x="Word", y="Count") +
  scale_fill_manual(values=c("green", "red")) +
  theme_minimal()
```