Advanced EDA of Video Text Dataset

Data Science Team

September 19, 2024

Dashboard

For seeing all the code live interactively,

Visit Dashboard

https://eda-analysis-iby-0.streamlit.app/

Contents

After all the preprocessing and cleaning of the data, I have created a new dataframe which contains the average scores of all the features for each student.
I have explained this process in main.pdf file.
So this is my new DATAFRAME , Let's call it FINAL_DF

Final Dataframe

id	avg_positive	avg_negative	avg_neutral	avg_confident
1	0.709199	0.141214	0.149586	0.733828
2	0.722006	0.107541	0.170453	0.684879

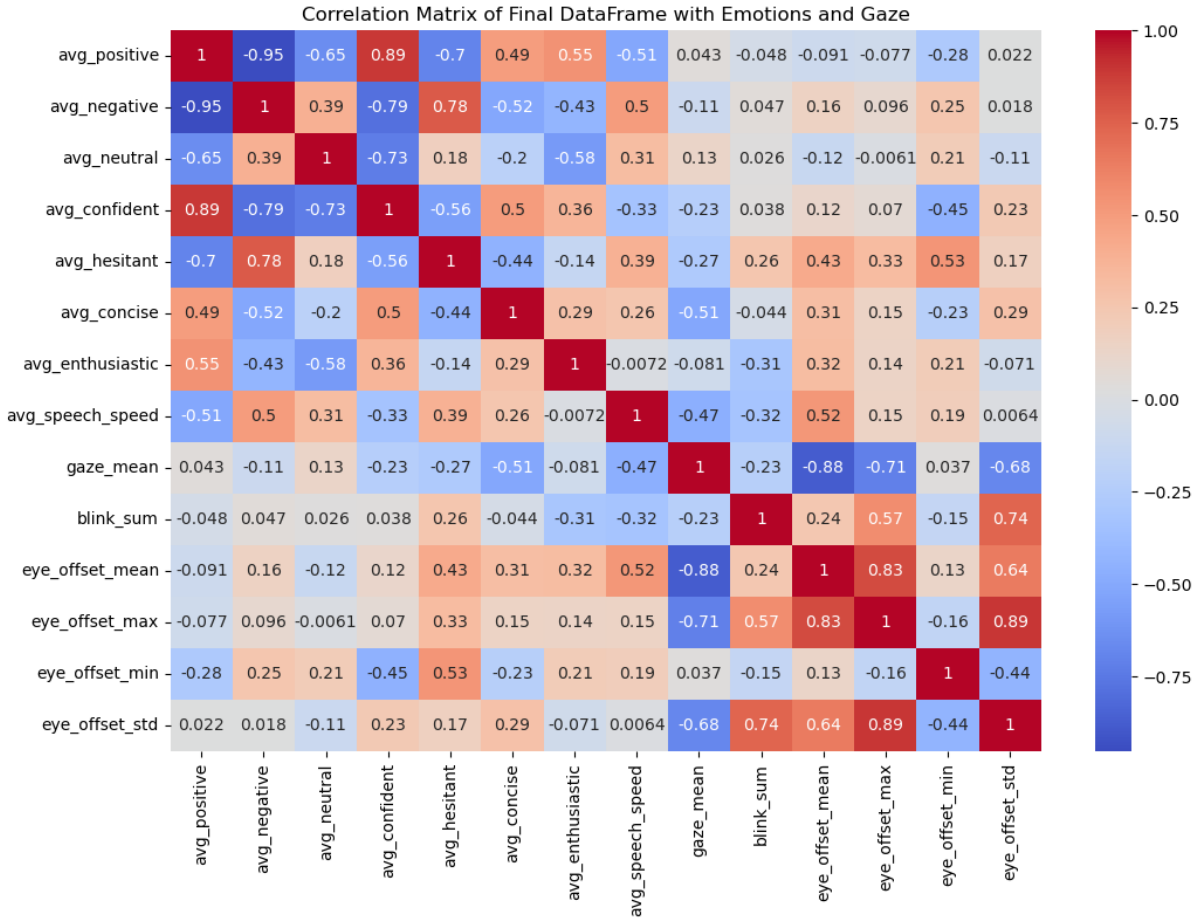
avg_hesitant	avg_concise	avg_enthusiastic	avg_speech_speed
0.485172	0.429418	0.466497	3.113771
0.436158	0.484221	0.516685	3.269092

dominant_emotion_top1	dominant_emotion_top2	gaze_score
neutral	fear	0.625000
happy	neutral	0.609195

blink_sum	eye_offset_mean	eye_offset_max	eye_offset_min
0.000000	15.801362	65.0276	-33.4655
4.597701	21.768546	67.6710	-15.2405

eye_offset_std	image_seq_count
17.858517	88
15.619435	87

2. BASIC Analysis



Insights from the correlation matrix

Focusing on this image, We can conclude :

- **avg_positve and avg_confident is highly correlated as red means they are correlated.**
- **avg_hesitant and avg_negative is also correlated as they are blue and their correlation score is 0.77(quite high) .**
- **avg_enthusiastic and avg_confident is also correlated though not as high as avg_positive and avg_confident.**
- **avg_negative and avg_confident is highly uncorrelated and their correlation score is -0.73(quite high) .**

In this way, we can see the correlation between the features like which features are dependent or which are not. From this we can if a student whose text content score is positive, then he/she is more likely more confident and enthusiastic in comparison to the student whose text content score is negative. This fact will help in further analysis.

Distribution plots were generated for various features to understand their spread and central tendencies. For example:

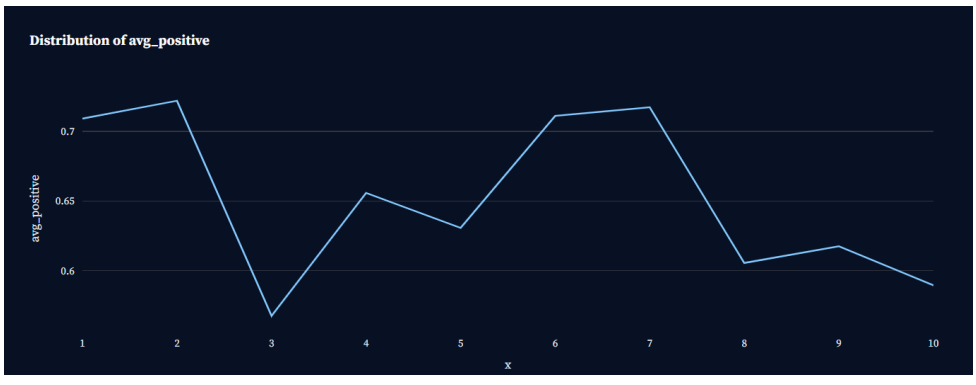


Figure 1: Distribution of avg_positive scores

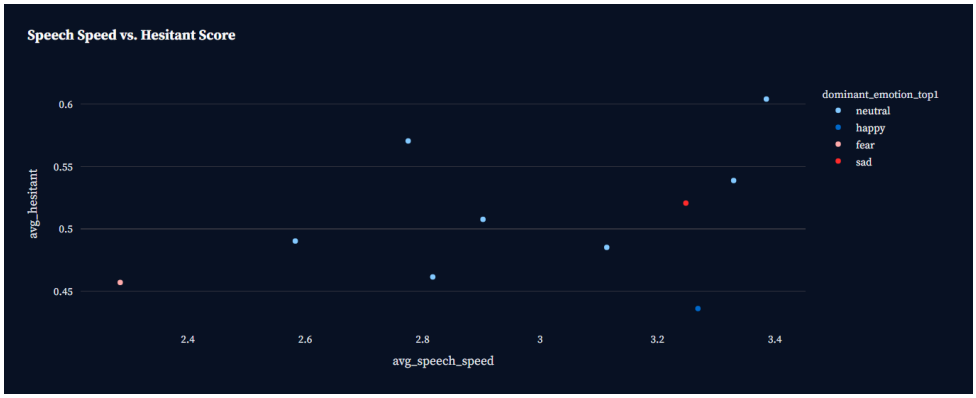
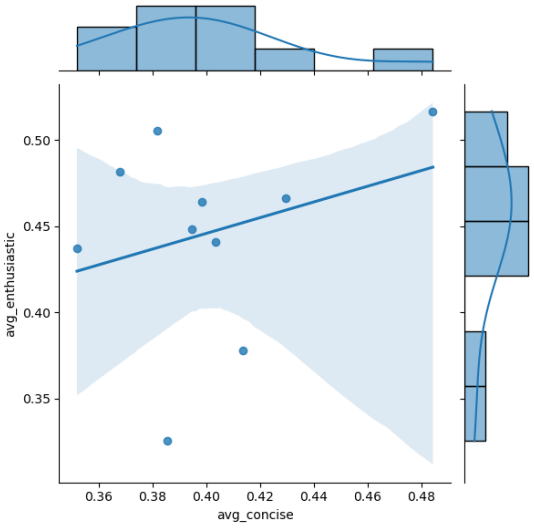


Figure 2: Speech Speed vs. Hesitant Scores

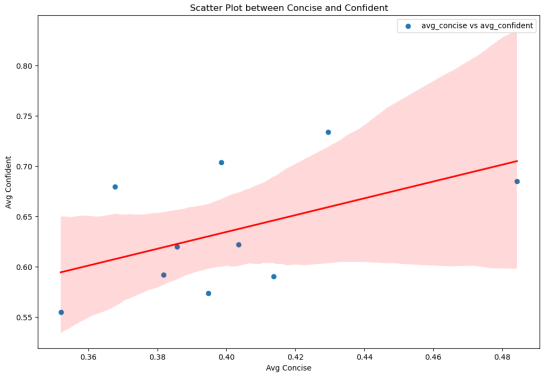
A.) Communication Skills Analysis

- **First**, I will investigate the relationship between **conciseness** and **enthusiasm** of the students.
  For this, I am plotting a joint plot between avg_concise and avg_enthusiastic
  From the plot, we can see that there is a positive correlation between the two features.

  This indicates that students who are more concise in their speech are also more enthusiastic.



Joint plot between avg_enthusiastic and avg_concise.

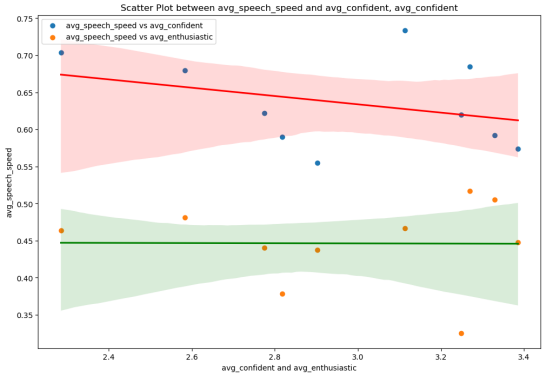


Scatter plot between avg_speech_speed and avg_hesitant.

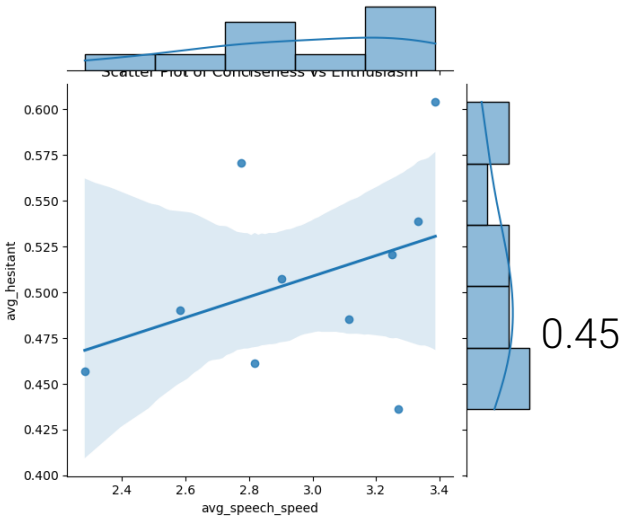
Comparison of conciseness vs enthusiasm and confidence.

- **Communication** skill is also dependent on the **speed of speech**. To analyze this, I am plotting a scatter plot between avg_speech_speed and avg_hesitant.

  The plot reveals a negative correlation between these two features, meaning that students who speak faster are less hesitant in their speech.



Scatter plot between avg_speech_speed and avg_confidence.



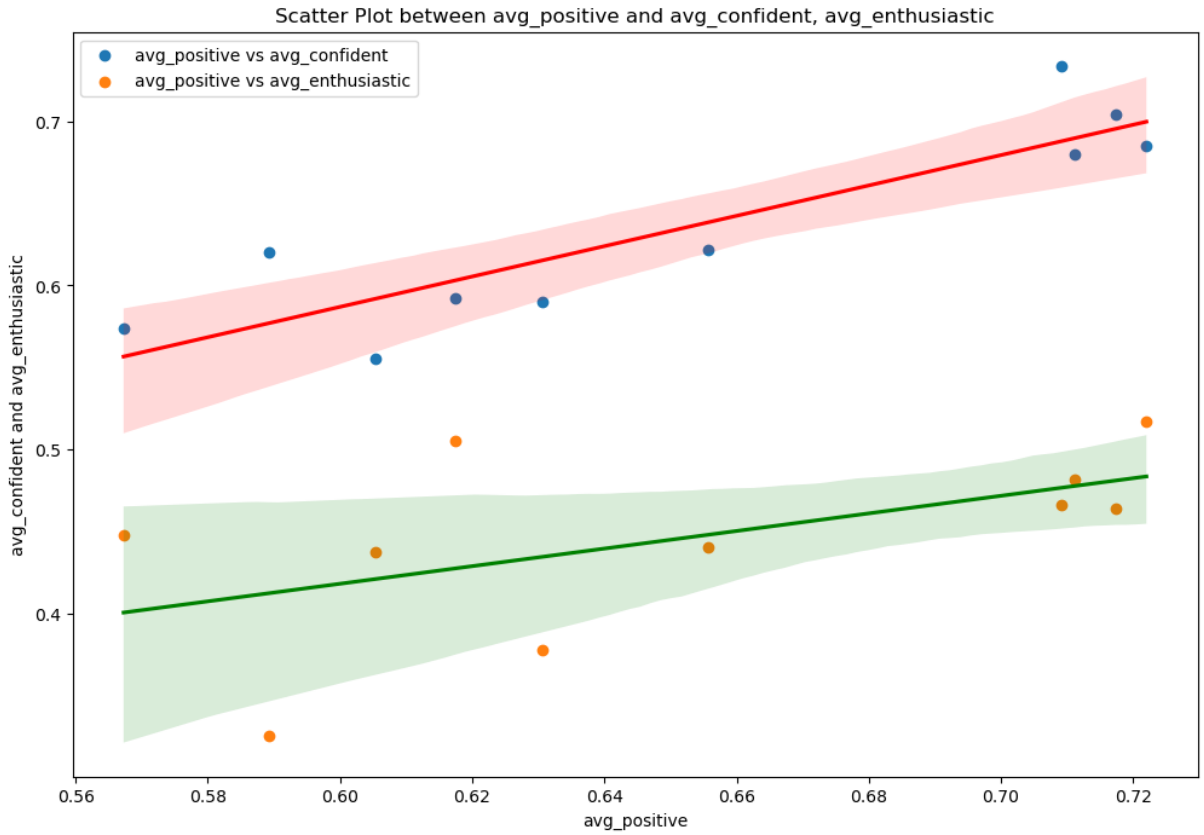
0.45

Scatter plot between avg_speech_speed and avg_hesitant.

- **Text** content scores (positive, negative, neutral) are also crucial in communication skills. Therefore, I will analyze the relationship

between **positivity** and **confidence** of the students.

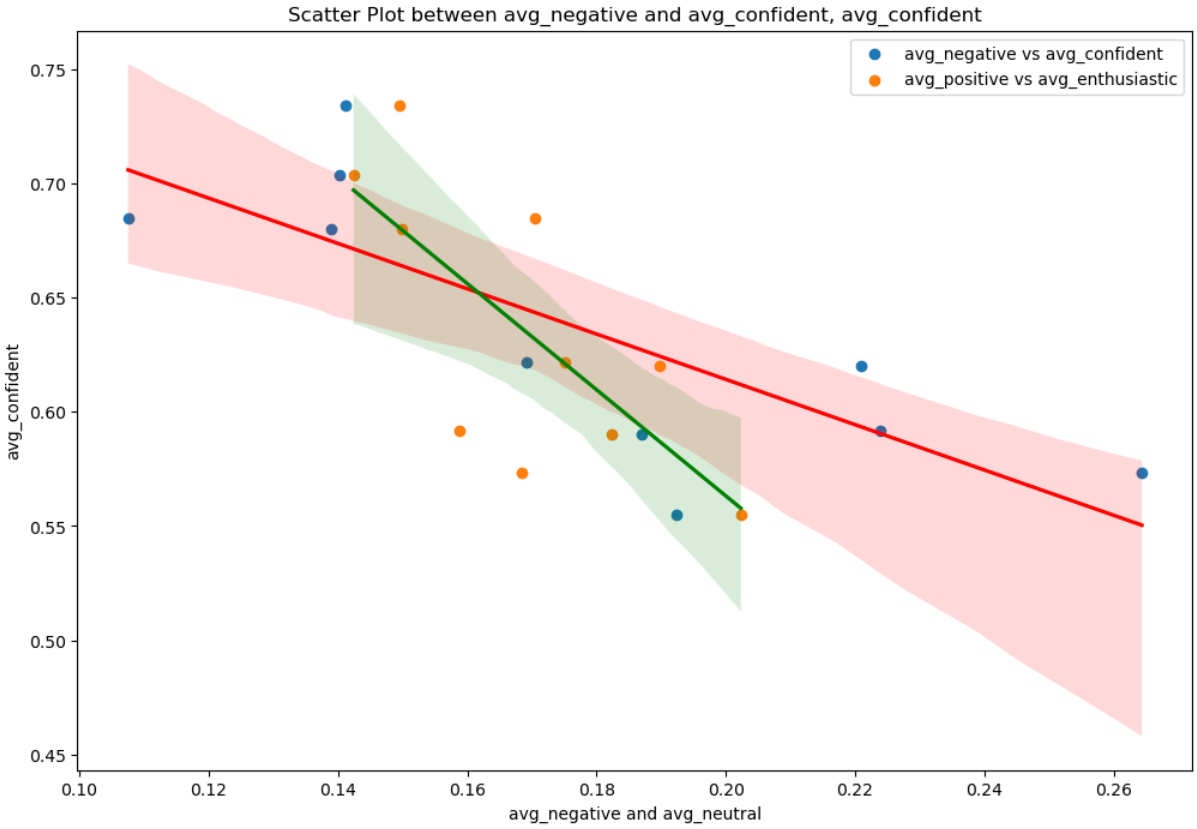
For this, I am plotting a scatter plot between avg_positive and avg_confident. The plot shows a positive correlation between these features, indicating that students with a more positive text content score are also more confident in their speech.

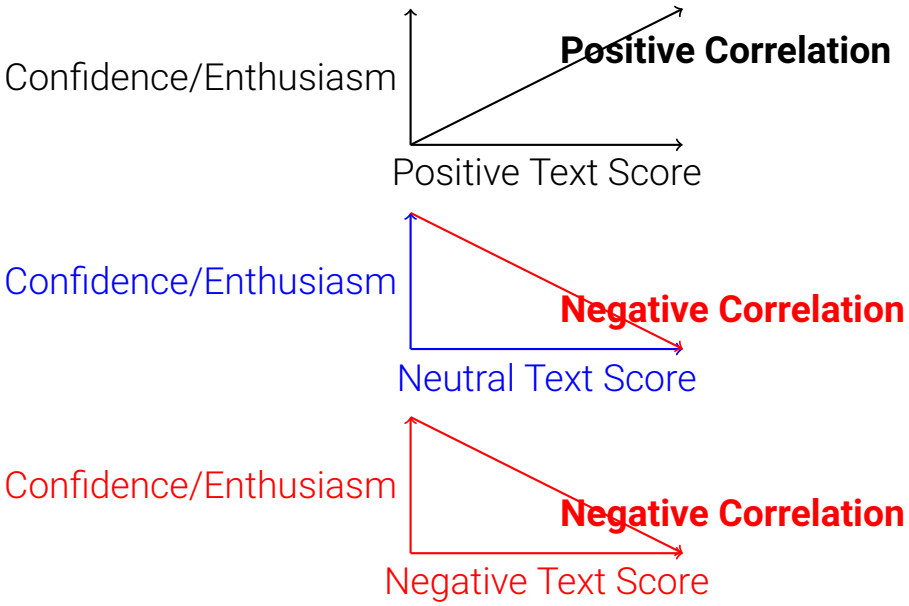


From this graph, we can clearly see the linear relation between avg_positive vs avg_confident and avg_positive vs avg_enthusiastic.

- **Additionally**, the relationship between avg_negative and avg_confident is also linear but in the opposite direction.

This implies that students with a more negative text content score tend to be less confident and enthusiastic.





Correlation Diagram for Text Scores with Confidence and Enthusiasm

- An interesting insight from the above graph is that avg_neutral is also linearly related to avg_confident and avg_enthusiastic, but in the opposite direction. This provides new insights into how neutrality in text content affects communication skills.

•

B.) Emotional State and Body Language Analysis

1.**Emotional Stability Analysis**
So, I have calculated the **emotional stability** of each student by analyzing the **variability** in their emotions throughout the video.

Emotional stability is a key factor in understanding how well students can manage their emotions during communication.
This is important because it helps us understand how consistent a student is in expressing different emotions.

> **Warning**
>
> Since there are 10 students and in this report, for a sample, I am showing the analysis for one student only. For other students' analyses, kindly visit the full report here.

There is slight moment in between where the student is showing fear and sadness.

This indicates that he is able to maintain a consistent emotional tone (which is neutral) during communication, which is a positive trait for

effective interaction.
`This student has shown a **high level of emotional stability** throughout the video, with minimal fluctuations in their emotional state.`

`There is slight moment in between where the student is showing fear and sadness.`

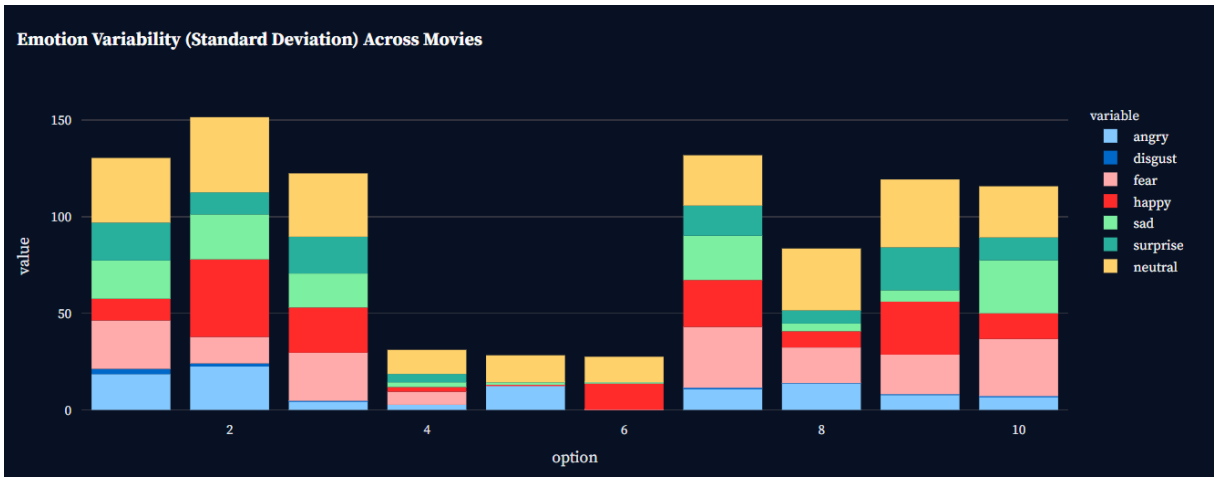
`This indicates that he is able to maintain a consistent emotional tone (which is neutral) during communication, which is a positive trait for effective interaction.`

2. **Emotion Variability Analysis**

Emotion variability is another important aspect of emotional intelligence, as it reflects how well students can adapt their emotions to different situations.

The amount of emotions a student is showing during their speech can be an indicator of their emotional intelligence.

If he shows fear or sadness for a long time, then it can be a sign of less communication skills and poor body language.



Like in this graph, we can see that except student 5 and 6, all stu-

dents are showing a good amount of variability in their emotions.

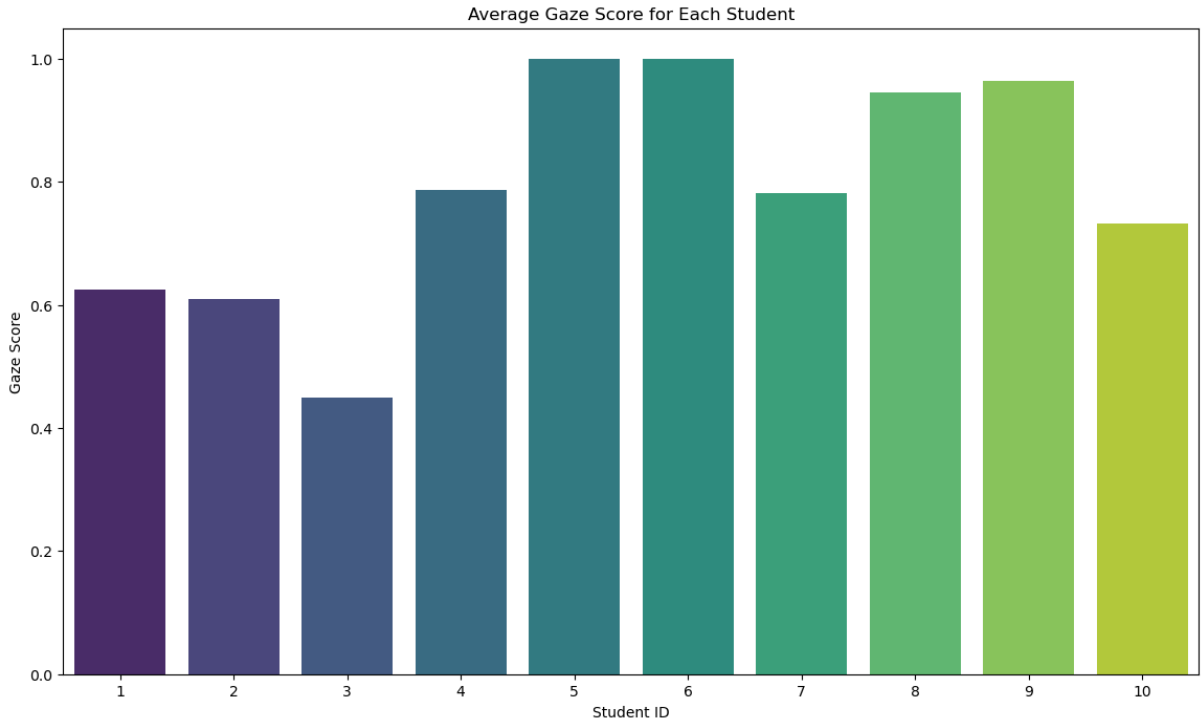
3. **Body Language Analysis**
So for this, I made a new dataframe named fiNAL-GAZE_DF which contains the gaze data of all the students.

It contains the following columns:

- `movie_id:` movie_id
- `Gaze_score:` The gaze score of the student: proportion of time the candidate spends looking at the camera..
- `blink_sum:` The blink sum of the student.
- `eye_offset_std:` The eye offset standard deviation of the student.

1. If the standard deviation (std) of the eye offset of a person in a video is too high, it suggests that the person's gaze is not stable or consistent across frames

i. **Gaze Analysis**
Gaze is an important aspect of body language that can reveal a lot about a student's focus and engagement during communication.



BASED ON THE ABOVE GRAPH, FOLLOWING OBSERVATIONS CAN BE MADE:

1. Highly Engaged (0.85 - 1.0): Students who maintained frequent or constant eye contact with the camera.
   In this category, Student 5.6,8,9 fall as their gaze score is above 0.85.

2. Moderately Engaged (0.7 - 0.85): Students who maintained moderate eye contact with the camera.
   In this category, Student 4,7,10 fall as their gaze score is between 0.5 to 0.85.

3. Low Engagement (Below 0.7): Students who maintained low eye contact with the camera.
   In this category, Student 1,2,3 fall as their gaze score is below 0.5.

ii. **Blink Analysis**

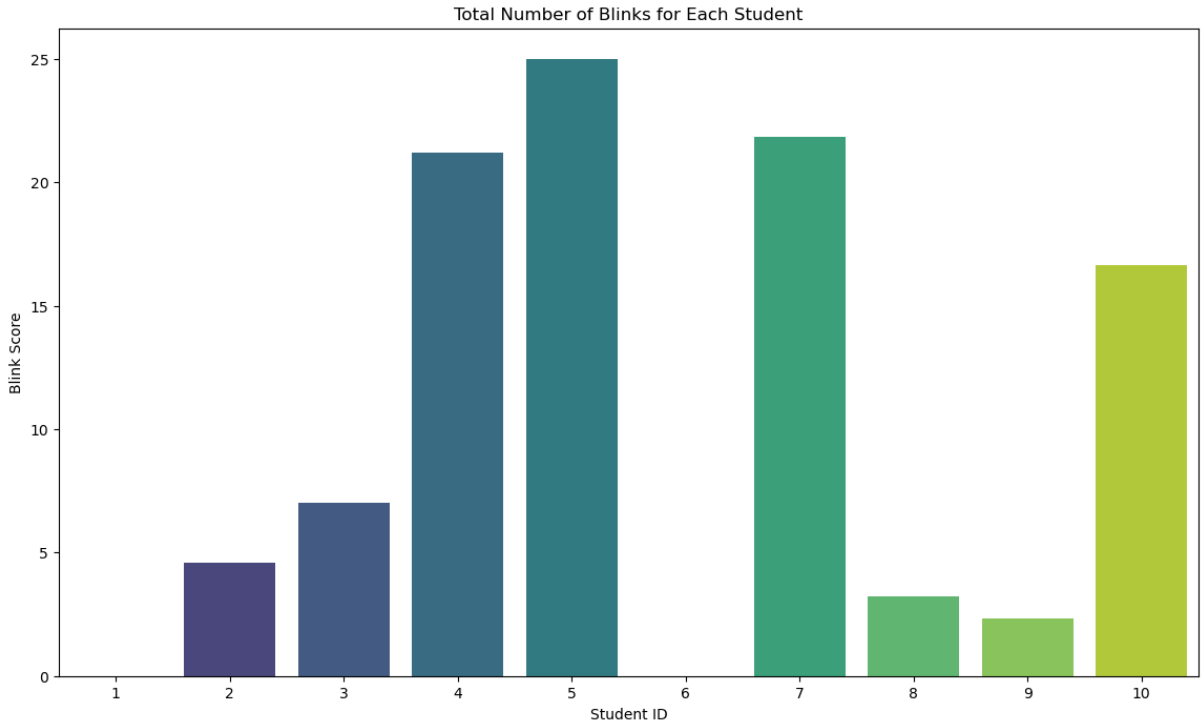
Blinking is another important aspect of body language that can indicate a student's level of comfort and confidence during communication.

If a student blinks too frequently, it may suggest nervousness or discomfort, while infrequent blinking may indicate confidence or focus.

Catch

Since the blinking rate will depend on the amount of time of video, i have to divided the blink$_sumwithtotal frames of the video to get the blink rate.$

Blink Rate = blink_sum / total_frames

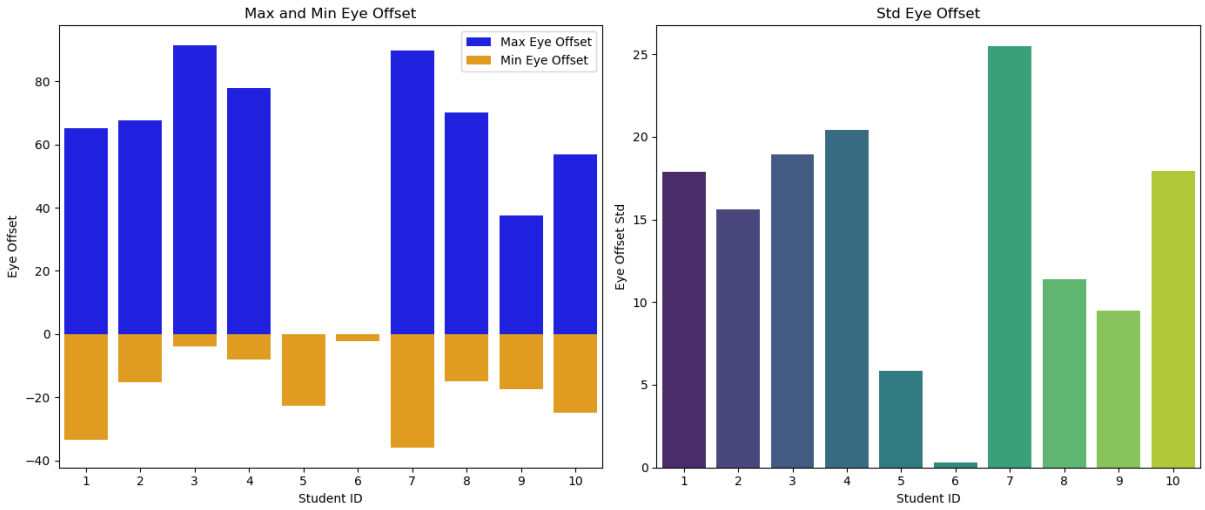


BASED ON THE ABOVE GRAPH, FOLLOWING OBSERVATIONS CAN BE MADE:

1. High Blink Rate: Students who blinked frequently during the video, i.e they are either nervour or feeling anxiety.
   In this category, Student 4,5,7,10 fall as their blink rate is .
2. Normal Blink Rate: Students who blinked at a moderate rate during the video.
   In this category, Student 1,2,3,5,6,8,9 fall as their blink rate is between 0.2 to 0.5.

iii. **Eye Offset Analysis**

Eye offset is a measure of how much a student's gaze deviates from the camera during communication.
So the higher the eye offset, the more the student's gaze is wandering away from the camera.



1. Max and Min Eye Offset (Left Plot):

This chart displays the maximum and minimum eye offsets for each student.
The blue bars represent the max positive deviation, and the orange bars represent the max negative deviation.

2. Eye Offset Standard Deviation (Right Plot):

This chart shows the standard deviation of eye offsets for each student.
The standard deviation indicates how much the student's eyes typically deviated from the mean eye position over the duration of the video.
BASED ON THE ABOVE GRAPH, FOLLOWING OBSERVATIONS CAN BE MADE:

1. `Highly Erratic Eye Movements:`

    - Student 7 shows the greatest overall deviation in both positive/negative offsets and in the standard deviation, indicating

highly variable and erratic eye movements.

- Students 3 and 4 also show significant deviation in eye movements, suggesting frequent or large fluctuations in where they were looking.

2. `Moderate Eye Movements:`

   - Students 1, 2, and 10 demonstrate moderate eye movement variability. They have noticeable deviations but are less erratic compared to students like 7 and 3.

3. `Stable Eye Movements:`

   - Students 5, 6, 8, and 9 exhibit the most stable eye movements, with minimal deviation from the mean eye position. This suggests that they maintained consistent eye contact with the camera throughout the video.

7. Candidate Evaluation Framework Explanation

This section provides a detailed explanation of the Python code used for evaluating candidates based on various metrics derived from their interview performance.

Overview

The framework consists of several key components:

- Score calculation functions for different aspects of the candidate's performance
- A function to categorize candidates based on their total score
- A main analysis function that applies the scoring framework to each candidate
- Data processing and result presentation

Score Calculation Functions

(b).1 Communication Score

$$\begin{aligned}\text{Communication Score} = \min(\text{avg\_positive} \times 10, 10) &+ \min(\text{avg\_neutral} \times 5, 5) \\ &+ \max(5 - |3 - \text{avg\_speech\_speed}| \times 2, 0) \\ &+ \text{avg\_concise} \times 5\end{aligned} \tag{1}$$

This function evaluates the candidate's communication skills based on:

- Positive and neutral language use
- Speech speed (with an ideal speed of 3)
- Conciseness

(b).2 Body Language Score

$$\text{Body Language Score} = gaze\_score*5+\max(5-\frac{eye\_offset\_std}{10},0)+\max(5-\frac{blink\_sum}{10},0) \tag{2}$$

This function assesses the candidate's body language, considering:

- Gaze direction
- Eye movement stability
- Blinking frequency

(b).3 Confidence and Enthusiasm Score

$$\text{Confidence \& Enthusiasm Score} = avg\_confident*10+avg\_enthusiastic*10 \tag{3}$$

This score is a direct measure of the candidate's perceived confidence and enthusiasm.

(b).4 Emotional Intelligence Score

$$\text{Emotional Intelligence Score} = \begin{cases} 5, & \text{if dominant\_emotion\_top1 in \{happy, neutral\}} \\ 5, & \text{if dominant\_emotion\_top2 in \{happy, neutral\}} \\ -5, & \text{if dominant\_emotion\_top1 in \{sad, angry, fear\}} \\ -5, & \text{if dominant\_emotion\_top2 in \{sad, angry, fear\}} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

This function evaluates the candidate's emotional range and appropriateness, rewarding positive emotions and penalizing negative ones.

(b).5 Composure Score

$$\text{Composure Score} = -(avg\_hesitant * 7) - (avg\_negative * 8) \tag{5}$$

This score reflects the candidate's ability to maintain composure, penalizing hesitancy and negative expressions.

Fitness Categorization

Candidates are categorized based on their total score:

$$\text{Fitness Category} = \begin{cases} \text{"Fit for role"}, & \text{if Total Score} \geq 35 \\ \text{"Okay Fit for role"}, & \text{if } 30 \leq \text{Total Score} < 35 \\ \text{"Not Fit"}, & \text{if Total Score} < 30 \end{cases} \tag{6}$$

Main Analysis Function

The `analyze_candidate` function applies all scoring functions to a candidate's data and returns a comprehensive evaluation, including:

- Individual scores for each evaluated aspect
- Total score
- Fitness category

Data Processing and Result Presentation

The framework processes the data as follows:

1. Applies the `analyze_candidate` function to each row of the input data
2. Combines the original data with the calculated results
3. Sorts candidates by their total score in descending order
4. Displays a summary table with candidate IDs, total scores, and fitness categories
5. Provides detailed feedback for each candidate, highlighting strengths and areas for improvement based on their scores in each category

Ranking of Students and Fit or Not

Candidate Evaluation Summary

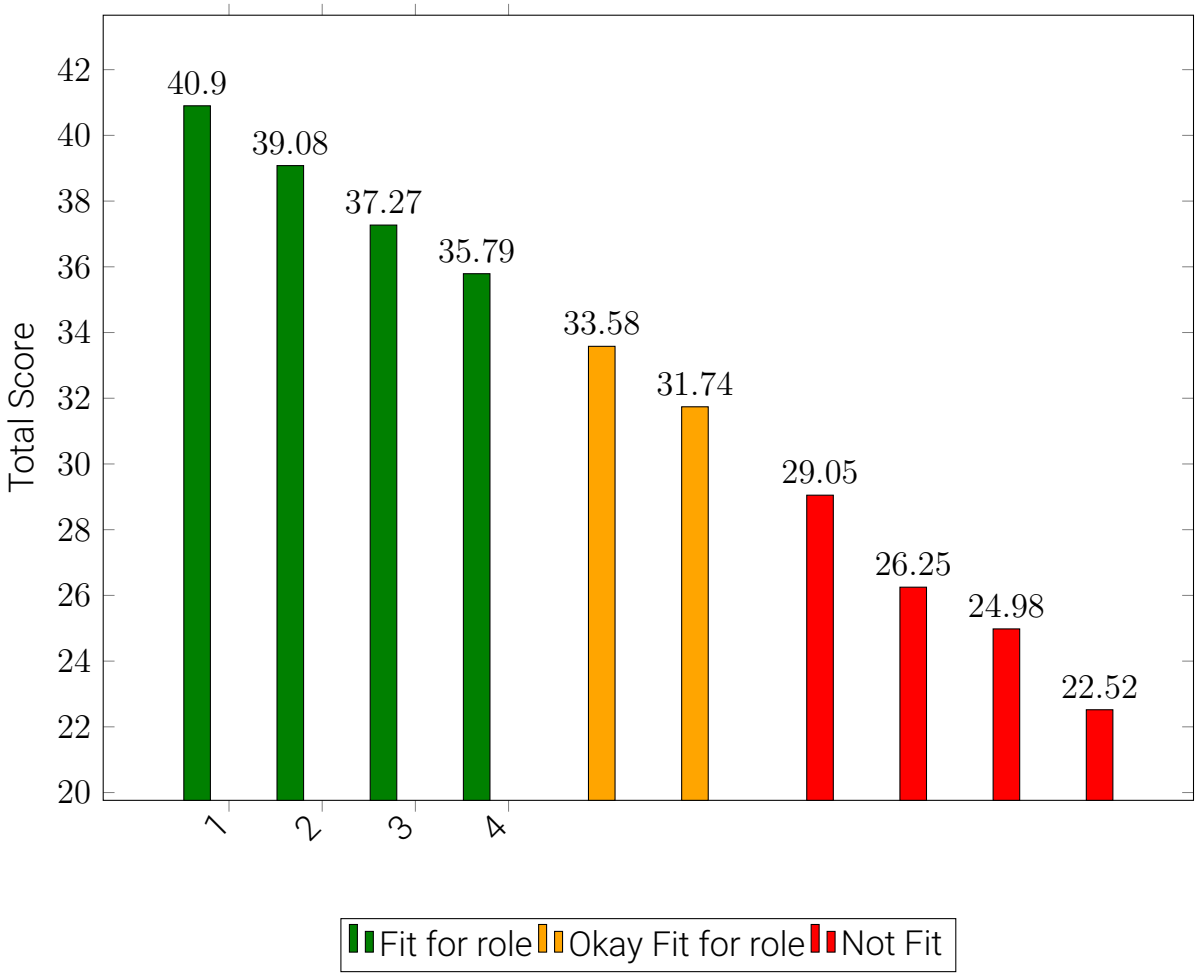


Figure 3: Candidate Total Scores

Candidate ID Color Code	id	Total Score	Fitness Category
b4b6b5a2-4203-41c2-b703-c424dae1fe2b	6	40.90	Fit for role
3d7cd21a-3170-4352-b499-24ea04eaf48c	2	39.08	Fit for role
deb4a835-b82f-4f3d-b2c4-77c66eca7752	9	37.27	Fit for role
80985461-c5d6-466f-a30a-4de2784ed0a3	5	35.79	Fit for role
e2aa9258-47a5-46ab-9c5c-283460f7a807	1	33.58	Okay for role
1c0c686b-3aae-4ac6-8625-3e86a7a0892f	8	31.74	Okay for role
62ea9b36-7860-4dc9-827c-600604286571	4	29.05	Not Fit
f299e1b2-7d92-4420-9c5a-d0d2590abdbe	3	26.25	Not Fit
d851fe95-3ead-47c1-88aa-d6fc453f7021	7	24.98	Not Fit
70a013ed-120a-41fa-bedd-75a5d15afb76	10	22.52	Not Fit

Table 1: Candidate Evaluation Overview

Detailed Candidate Evaluations

Fit for Role Candidates

(a).1 Candidate ID: 92016995-e455-4651-9f6e-fbca0d423f21

- **Total Score:** 40.90
- **Fitness Category:** Fit for role
- **Areas for Improvement:**
  - Enhance communication skills
  - Work on body language
  - Boost confidence and enthusiasm
  - Develop emotional intelligence
  - Improve composure under pressure

(a).2 Candidate ID: baa26895-85b2-465b-a972-649b41d9870e

- **Total Score:** 39.08
- **Fitness Category:** Fit for role
- **Areas for Improvement:**
  - Enhance communication skills
  - Work on body language
  - Boost confidence and enthusiasm
  - Develop emotional intelligence
  - Improve composure under pressure

(a).3 Candidate ID: dfb0d746-609f-4dac-8e1d-c0325fb64394

- **Total Score:** 37.27
- **Fitness Category:** Fit for role
- **Areas for Improvement:**
  - Enhance communication skills
  - Work on body language
  - Boost confidence and enthusiasm
  - Develop emotional intelligence
  - Improve composure under pressure

(a).4 Candidate ID: 9c350343-e895-49df-af90-d50b91d19d3e

- **Total Score:** 35.79
- **Fitness Category:** Fit for role
- **Areas for Improvement:**
  - Enhance communication skills
  - Work on body language
  - Boost confidence and enthusiasm
  - Develop emotional intelligence
  - Improve composure under pressure

Okay Fit for Role Candidates

(b).1 Candidate ID: 93663f94-bf0a-4ce8-a29a-a5236cc7fe6a

- **Total Score:** 33.58
- **Fitness Category:** Okay Fit for role
- **Areas for Improvement:**
  - Enhance communication skills
  - Work on body language
  - Boost confidence and enthusiasm
  - Develop emotional intelligence
  - Improve composure under pressure

(b).2 Candidate ID: 813af424-a584-4417-b7ee-0d4c705e83c9

- **Total Score:** 31.74
- **Fitness Category:** Okay Fit for role
- **Areas for Improvement:**
  - Enhance communication skills
  - Work on body language
  - Boost confidence and enthusiasm
  - Develop emotional intelligence
  - Improve composure under pressure

Not Fit Candidates

(c).1 Candidate ID: 6b0386fc-41de-4196-b0d6-3d0b815c2dbc

- **Total Score:** 29.05
- **Fitness Category:** Not Fit
- **Areas for Improvement:**
  - Enhance communication skills
  - Work on body language
  - Boost confidence and enthusiasm
  - Develop emotional intelligence
  - Improve composure under pressure

(c).2 Candidate ID: d0b9170b-98b9-48e1-a1b2-1d661bb0d853

- **Total Score:** 26.25
- **Fitness Category:** Not Fit
- **Areas for Improvement:**
  - Enhance communication skills
  - Work on body language
  - Boost confidence and enthusiasm
  - Develop emotional intelligence
  - Improve composure under pressure

(c).3 Candidate ID: 6539370c-256e-4ed2-9d00-1be1f051163f

- **Total Score:** 24.98
- **Fitness Category:** Not Fit
- **Areas for Improvement:**
  - Enhance communication skills
  - Work on body language
  - Boost confidence and enthusiasm
  - Develop emotional intelligence
  - Improve composure under pressure

(c).4 Candidate ID: 83c20b83-7881-499d-a40d-cc06b65869f8

- **Total Score:** 22.52
- **Fitness Category:** Not Fit
- **Areas for Improvement:**
  - Enhance communication skills
  - Work on body language
  - Boost confidence and enthusiasm
  - Develop emotional intelligence
  - Improve composure under pressure