Exploratory Data Analysis Report

Saurabh Kumar Sah

21110188

IIT Gandhinagar

Task: To generate valuable and actionable insights from the dataset of 10 candidates containing emotion scores, transcript scores, and the corresponding transcripts extracted from their introduction videos.

27/09/2023

Table of Contents:

Executive Summary

This report consists of the Exploratory Data analysis of 10 candidates based on the data provided by the company. This report consists of all the steps involved in the process of assigning rank to each candidate. Based on the number of job openings, the company can pick as many candidates based on their rank calculated in the report.

In this project, I calculated the multiple ranks based on the data provided by the company, and then I calculated the final rank using the concept of weighted mean. This is because every factor does not account for equal values and equal importance. I have clearly mentioned all the weights given to factors in the report.

I have used various tools like Jupyter Notebook, Microsoft Excel, Google Sheets and Python programming language to perform the analysis. In Python, I have used pandas for calculations, matplot for plotting graphs, and the Excel writer function of pandas to perform certain calculations in an Excel file.

Our EDA report used bar graphs to efficiently convey categorical data patterns. Their simplicity aids in presenting data distributions and comparisons with clarity.

The report includes all the links for jupyter notebooks and Excel/CSV files.

Chat GPT was used in this analysis; its documentation is provided in a separate report.

Introduction:

“I am beside you” corporation wants to hire summer interns. For this purpose, they want me to analyse the emotional and transcript data for 10 candidates. This report contains the final report of the same analysis. They have performed video analysis of candidates for the same purpose.

Datasets used:

1. **Emotion Scores:** This dataset contains the candidate's emotions throughout the video.
2. **Transcript Scores:** This dataset contains scores extracted from the transcripts throughout the video.
3. **Transcript Text**: This column contains the actual text of the transcript of the video.

The source of all the datasets is the corporation itself.

Objectives:

1. Evaluate candidate suitability for recruitment based on emotion and transcript scores, providing clear reasons for hiring decisions.
2. Analyze communication skills and pinpoint areas of expertise using available data.
3. Analyze candidate data to identify communication skills, areas of expertise, and any additional insights crucial for making informed hiring decisions.

Data Description:

All the data except transcription text was in .csv format.

1. **Transcript Scores:**

   The following table describes the columns with their data types.

Column	Data types	Description
id	object	Unique credentials for each Candidate
seek	float64	-
start	float64	start time of the text
end	float64	end time of the text
text	object	Transcript spoken from start to end duration
tokens	object	-
temperature	float64	-
avg_logprob	float64	-
compression_ratio	float64	-
no_speech_prob	float64	-
positive	float64	positive score
negative	float64	negative score
neutral	float64	neutral score
confident	float64	confidence score
hesitant	float64	hesitance score
concise	float64	concise score
enthusiastic	float64	enthusiasm score
speech_speed	float64	Speed of the speech spoken

2. **Emotional Scores:**
   The following table describes the columns with their data types.

Column	Data types	Description
movie_id	object	Unique ID for video
image_seq	float64	Number of images
angry	float64	angry emotion score
disgust	float64	disgust emotion score
fear	float64	fear emotion score
happy	float64	happy emotion score
sad	float64	sad emotion score
surprise	float64	surprise emotion score
neutral	float64	neutral emotion score
dominant_emotion	object	dominant emotion, among other emotions

3. **Gaze:**
   The following table describes the columns with their data types.

Column	Data types	Description
movie_id	object	Unique ID for video
image_seq	float64	Number of images
gaze	float64	Candidate is looking at the camera or not. 1 for looking and 0 for not looking
blink	float64	Eye blink (1 -> blink, 0 -> No blink)
eye_offset	float64	Deviation of eye from the camera

4. **Metadata:**
   The following table describes the columns' data types.

Column	Data types	Description
movie_id	object	Unique ID for video
image_seq	float64	Number of images
participant_id	object	ID for participant
elapsed_time	float64	Timestamp in seconds(9 means at 9th sec)
upload_time	object	-
distance	float64	-

5. **Transcript text:**
   This contains the original transcript of the candidates, but their names are changed. The data of the transcript text was provided in the transcript score file.

Data Cleaning:

Although the data was clean and ready to be analysed, there were no missing values, duplicate records or any other anomalies. One strange thing that I encountered during the data cleaning process is that for candidate 1, the metadata under emotional data had more than one participent_id. This might be due to the presence of any other person in the frame for that image, and I had also taken account of those participant IDs. I left it to the higher authorities of the company whether to select Candidate 1 or not.

Analysis:

Step 1:

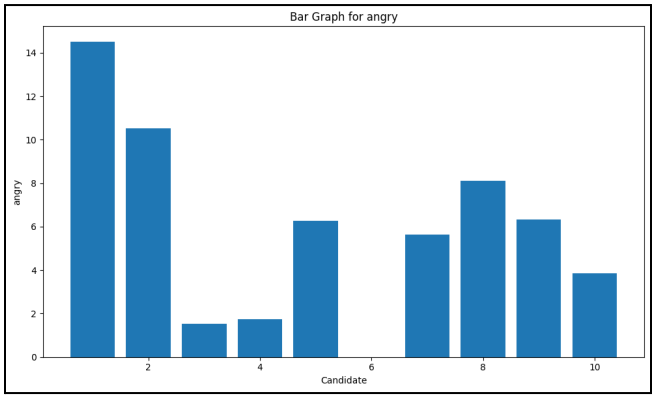
In our Exploratory Data Analysis (EDA), it's important to acknowledge and address the variation in the number of data points among candidates within our datasets. The unequal distribution of data points could bias our analyses. To mitigate this, we've taken a crucial first step by calculating the average across all data points for each candidate.

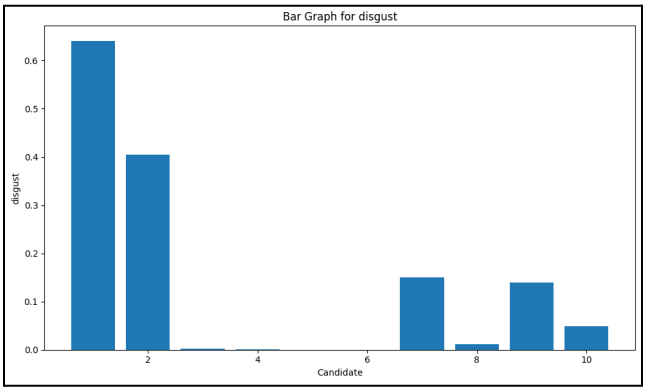
This approach ensures that we derive meaningful insights while accommodating the inherent data discrepancies. By using the average, we provide a fair representation of each candidate's performance, irrespective of the number of data points available. This enables us to make more balanced comparisons and draw reliable conclusions.

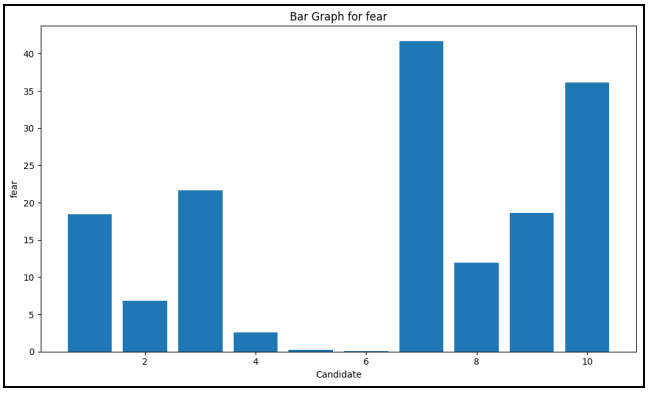
Step 2:

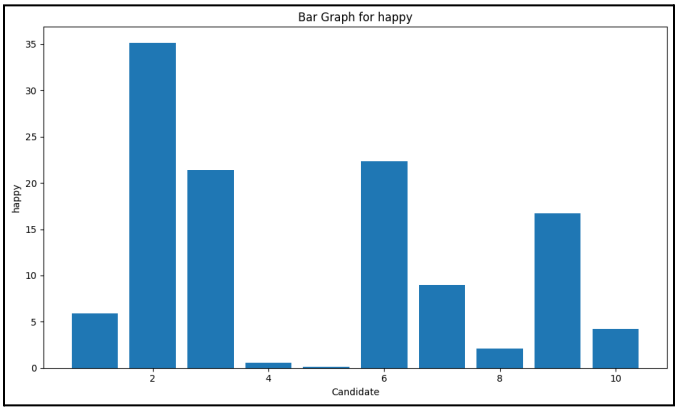
I compared the individual parameters using bar plots.

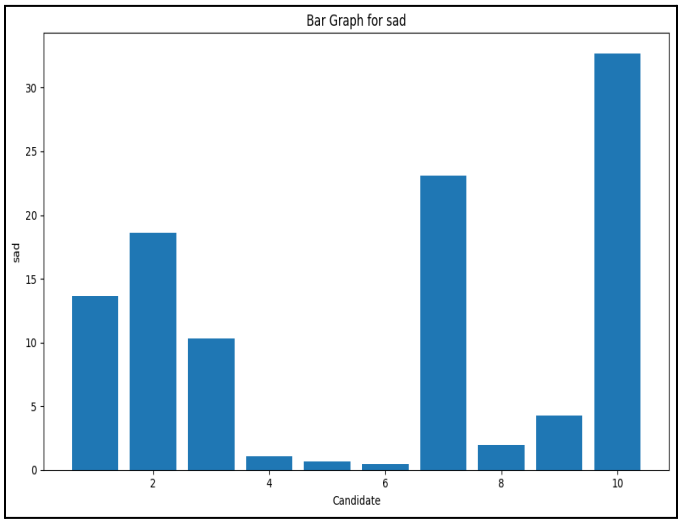
a) Emotion Score

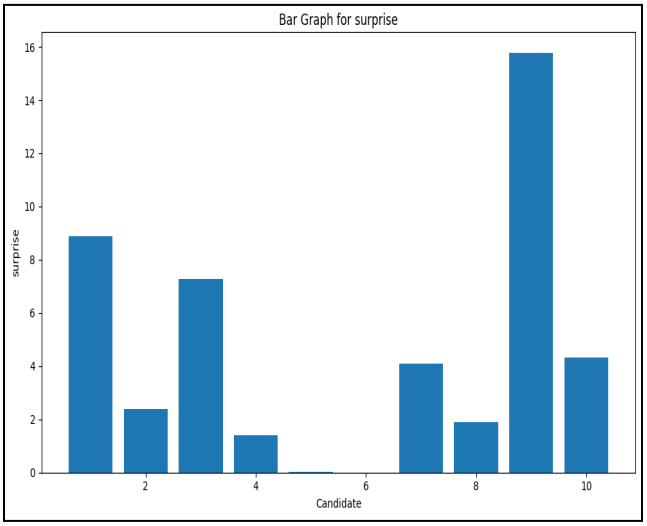


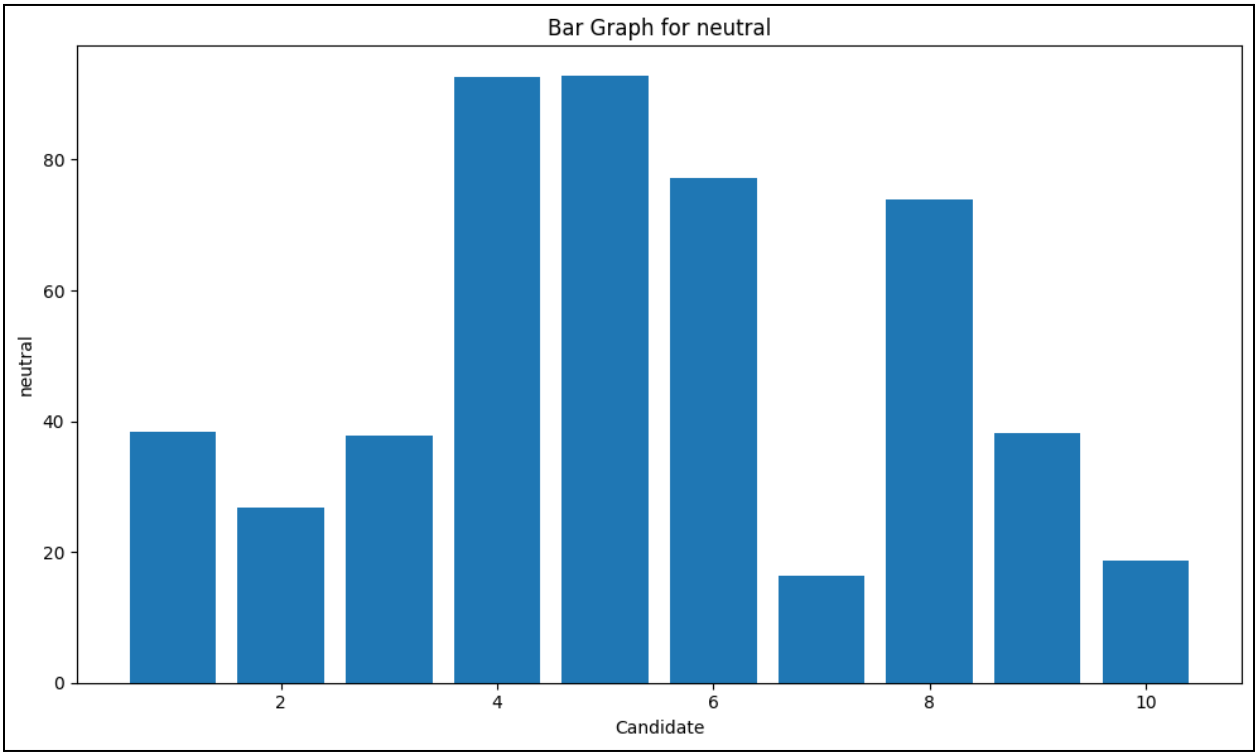




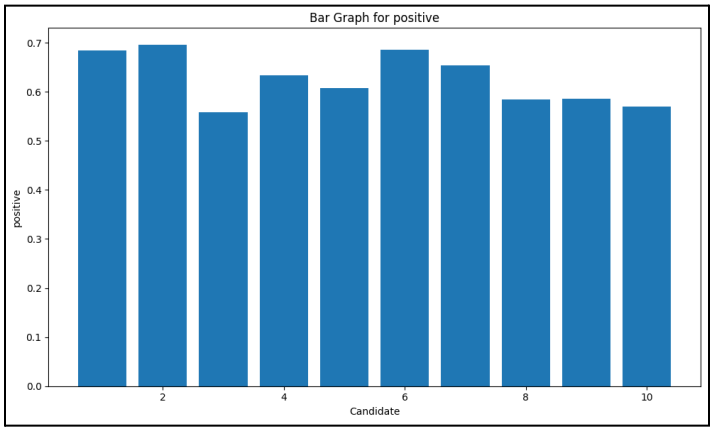


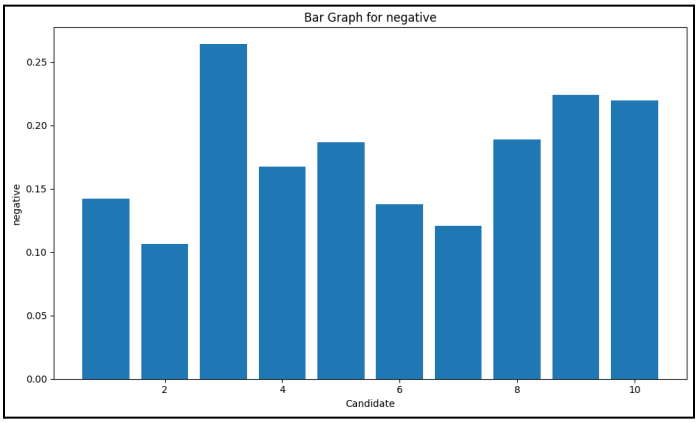

Bar Graph for sad
sad
Candidate
0
5
10
15
20
25
30
2
4
6
8
10

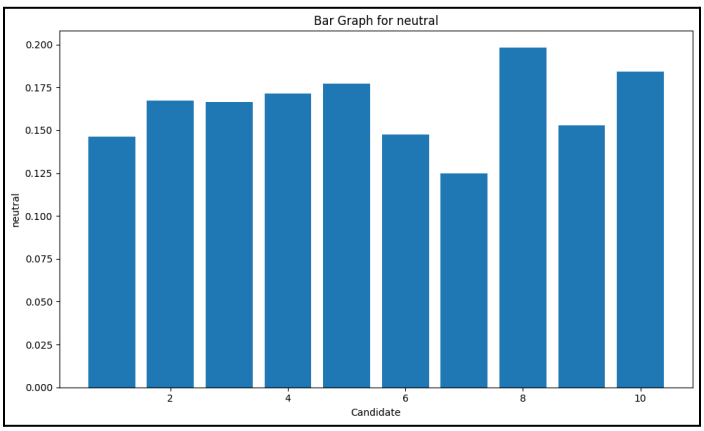

Bar Graph for surprise
surprise
Candidate
0
2
4
6
8
10
12
14
16
2
4
6
8
10

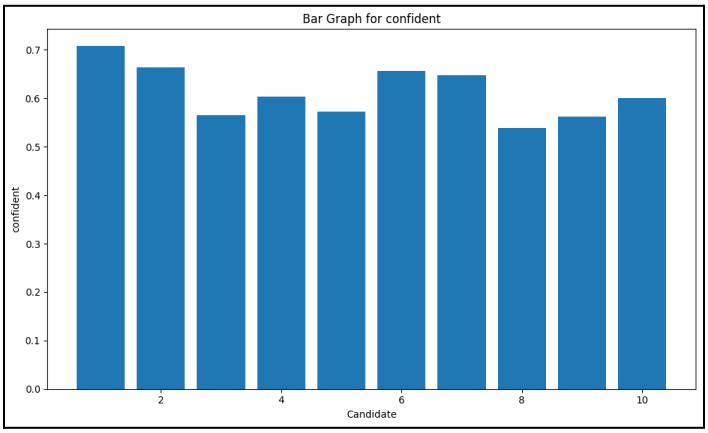

Bar Graph for neutral
neutral
Candidate
0
20
40
60
80
2
4
6
8
10

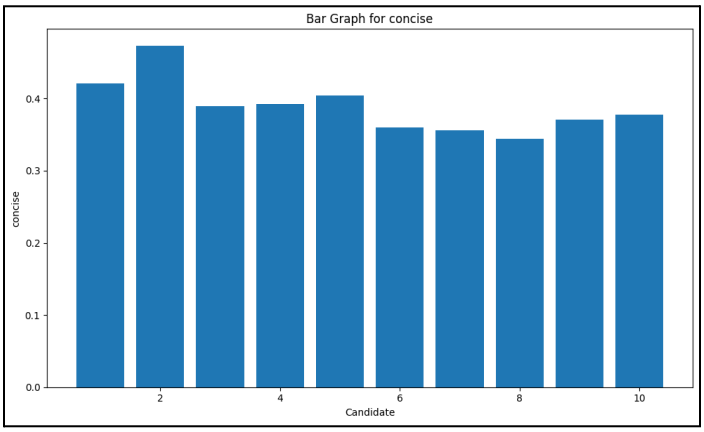
b) Transcript Score

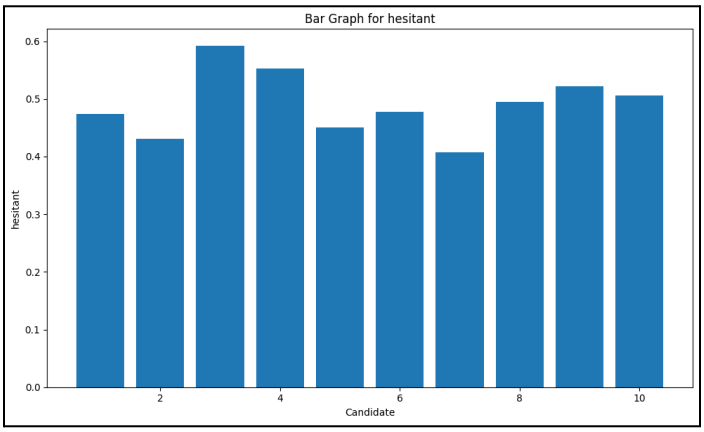


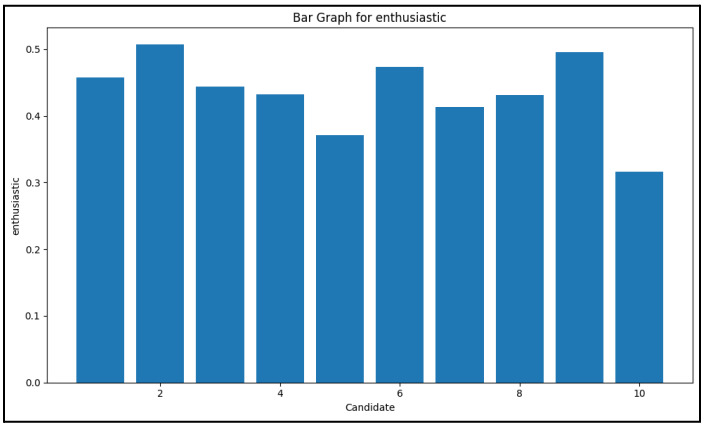


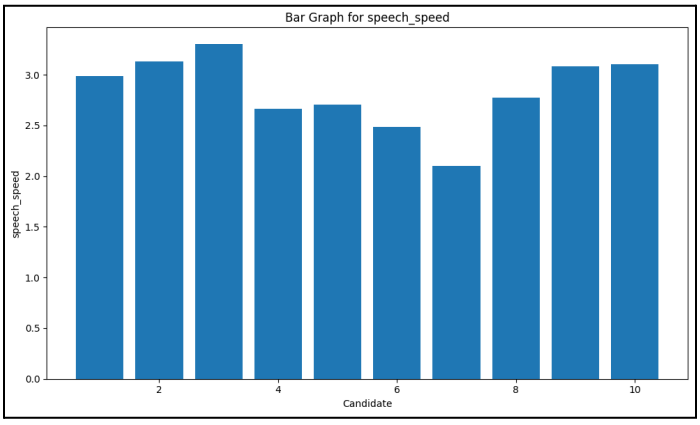












In EDA, I generate bar graphs to explore individual parameters. However, these visualisations yield only some meaningful insights. This initial lack of clarity may be attributed to various factors, such as the complexity of data, factors affecting individuals' performance simultaneously, insufficient contextual information, or the influence of unaccounted variables such as gaze and eye offset.

While not immediately enlightening, these bar graphs remain a valuable starting point in our analysis. They signal the need for a more in-depth investigation, including potential data aggregation, correlation assessments, or subgroup analyses.

Step 3:

In our exploratory data analysis (EDA), we used the concept of assigning weights to multiple parameters as part of our analytical process. This weighting approach is a deliberate strategy employed to ensure a more comprehensive and balanced data assessment.

Justification for this approach lies in recognising that not all parameters carry equal significance in our analysis. Some parameters may inherently possess greater relevance or impact on our research objectives or the overall context of the study. Assigning weights enables us to quantitatively account for these variations in importance.

Furthermore, the weighting strategy is rooted in the pursuit of more accurate insights. By considering the relative influence of each parameter, we aim to reduce the potential for skewed or misleading results. This approach aligns with best practices in data analysis, where it is common to adjust for variable importance to reflect real-world scenarios accurately.

It is important to emphasise that these weights have been assigned thoughtfully to EDA goals and data characteristics. This approach enhances the rigour and credibility of our EDA, ensuring that our findings are driven by a nuanced understanding of the parameters' significance.

For emotional parameters, the weights are as follows:

S. No.	Parameter	Weight
1	angry	-0.4
2	disgust	-0.2
3	fear	0.2
4	happy	0.5
5	sad	0.2
6	surprise	0.3
7	neutral	0.4

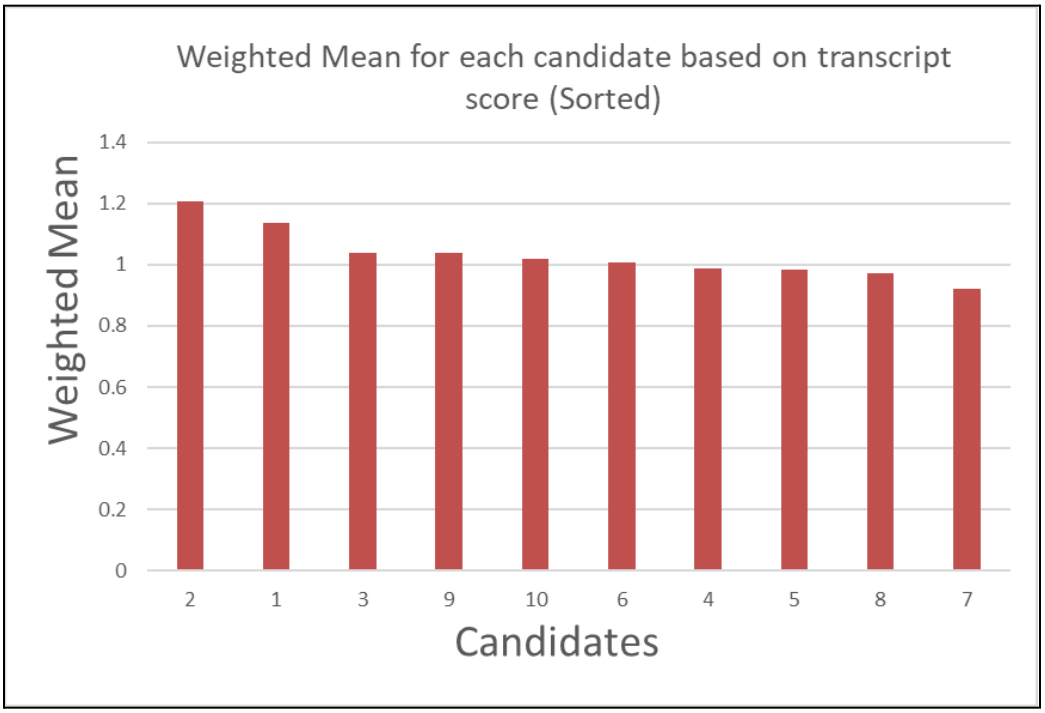
For transcript parameters, the weights are as follows:

S. No.	Parameter	Weight
1	positive	0.3
2	negative	-0.3
3	neutral	0.2
4	confident	0.35
5	hesitant	-0.25
6	concise	0.25
7	enthusiastic	0.25
8	speech_speed	0.2

After calculating the weighted mean for each candidate for transcript score and emotional score, we get the following data:

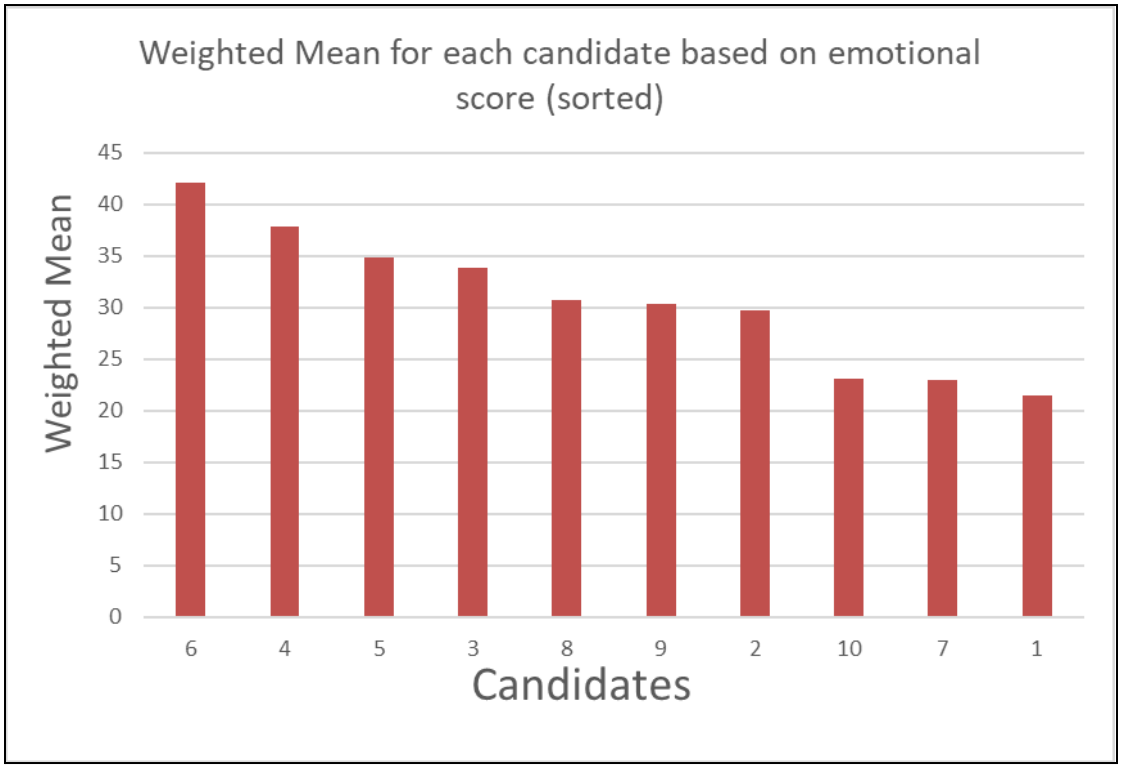
For transcript data:

Rank	Candidate	Weighted Mean
1	2	1.206448
2	1	1.138429
3	3	1.03991
4	9	1.03895
5	10	1.020384
6	6	1.009068
7	4	0.985797
8	5	0.984811
9	8	0.972447
10	7	0.921938



For emotional data:

Rank	Candidate	Weighted Mean
1	6	42.138191
2	4	37.797335
3	5	34.828597
4	3	33.804156
5	8	30.743969
6	9	30.364253
7	2	29.751418
8	10	23.114012
9	7	22.936152
10	1	21.440238



Note that the candidates are sorted in decreasing order of their weighted means. In other words, I had assigned ranks to each candidate separately for transcript and emotional score.

Step 4:
We have incorporated a data-driven prioritisation strategy based on **eye offset** values to enhance the decision-making process for the gaze dataset. This approach assigns scores to candidates according to the proximity of their eye offset values to a predetermined threshold, providing valuable insights for candidate selection.

The justification for this inclusion lies in its potential to optimise resource allocation. By assigning higher scores to candidates with eye offsets close to zero, we acknowledge their alignment with our preferred criteria—candidates looking neither above nor below the camera.

Furthermore, this data-driven prioritisation approach promotes fairness and objectivity by mitigating potential biases associated with manual selection. It maximises the likelihood of selecting candidates with optimal eye offset values, ultimately contributing to more accurate and efficient decision-making in our selection process.

The following table depicts the rank of candidates based on the gaze data.

Rank	Candidate	Average Eye Offset	Score
1	6	1.707193	2
2	8	6.56464	1.6
3	9	8.58629	1.6
4	7	9.456552	1.6
5	10	11.49859	1.2
6	4	12.49248	1.2
7	5	15.80136	0.8
8	1	15.80263	0.8
9	2	21.76855	0.4
10	3	30.13721	0

Step 5:
Final rank distribution for all the candidates:

For this, I have assigned weights to both the transcript score and emotional score as follows:

Emotions score Weight	40%
Transcript score Weight	60%

The decision to allocate a higher weight to the Transcript Dataset reflects our primary research focus on understanding students' performance and communication from the introduction video. The Transcript Dataset provides valuable information about the content of the interviews, including the clarity of their responses, the depth of their knowledge, and their ability to effectively convey their thoughts.

While emotions are important indicators of the interview experience, they are considered secondary factors in our analysis. By assigning a lower weight to the Emotions Dataset, we acknowledge its significance in capturing the emotional context of videos without overshadowing the primary goal.

After calculating scores from the above two datasets for each candidate, we get the following ranks:

Rank	Candidate	Final Score
1	6	17.4607172
2	4	15.7104122
3	5	14.5223254
4	3	14.1456084
5	8	12.8810558
6	9	12.7690712
7	2	12.624436
8	10	9.8578352
9	7	9.7276236
10	1	9.2591526

Finally, adding the gaze score for each candidate, we get the final rank as

Rank	Candidate	Final Score
1	6	19.46072
2	4	16.91041
3	5	15.32233
4	8	14.48106
5	9	14.36907
6	3	14.14561
7	2	13.02444
8	7	11.32762
9	10	11.05784
10	1	10.05915

Code or scripts

Links for the jupyter notebooks and Excel file used for performing EDA are

1. Jupyter Notebooks
    - Emotional Dasaet
    - Transcript Dataset
2. Excel files
    - Emotional Dataset
    - Gaze
    - Transcript Dataset
    - Final Score

Findings and Insights

1. Overall, candidate 6 is the best candidate for the job role.
2. The top three candidates for the job role are 6, 4, and 5 respectively.
3. Based on emotional data alone, candidate 6 is again the best candidate for the job.
4. Based on transcript data alone, the candidate's best fit for the job role is 2.

Other notable analysis findings:

a) Based on emotional parameters

- Neutral is the highest shown emotion by almost all candidates.
- Candidate 6, who is the overall best fit, has higher values for happy and neutral and lower values for disgust, fear, and sad.
- Candidate 1, who is the least fit for the role, has the highest value for disgust.
- Candidate 7 has the least value for neutral.

b) Based on transcript parameter:

- Candidate 2 has the most concise transcript.
- Candidate 3 is most hesitant.
- Candidate 2 is the most enthusiastic, and candidate 10 is the least.
- Candidate 7 has the least speech speed, and candidate 2 has the most.

Conclusion

Ranking candidates allows for easy individual comparison. It provides a clear hierarchy of candidate suitability, making it simpler for decision-makers to identify top-performing candidates quickly and efficiently.

The final scores can be tailored to align with the specific requirements of different job roles within the company. This customisation ensures that candidates selected have attributes and qualifications that closely match the job's demands.

Ranking candidates based on their performance and final scores is a crucial step in the recruitment process that offers several significant advantages for organisations. This structured approach simplifies the evaluation process and enhances the overall decision-making efficiency.

First and foremost, candidate ranking provides a concise and transparent representation of each candidate's suitability for a particular job role. It distils the extensive pool of applicants into a clear hierarchy, allowing decision-makers to identify the top-performing candidates quickly and efficiently. This is especially valuable when there are numerous applicants for a limited number of positions. With ranking, the process of manually comparing and contrasting individual qualifications and experiences could be more efficient and prone to human error.

For example,
The top 3 candidates can be selected as advanced data scientists, while the next three as data scientists. Following the same hierarchy for other candidates, too.

Appendices:

I extensively used ChatGPT as a supplementary resource during this project, which greatly enriched the overall project experience. This utilisation of ChatGPT served as an invaluable tool, providing a wide range of benefits and enhancing various aspects of the project.

One of the key advantages of leveraging ChatGPT was the ability to access additional information swiftly and effectively. When encountering unfamiliar concepts or seeking clarification on specific topics, I could promptly request explanations or definitions from ChatGPT. This not only saved time but also ensured that I had a clear understanding of the subject matter, contributing to the accuracy of my work.

Furthermore, ChatGPT played a pivotal role in expediting certain technical tasks. I could request code snippets and programming solutions tailored to the project's requirements. This accelerated the development process and facilitated the implementation of complex algorithms and data manipulation tasks. ChatGPT's ability to generate code snippets based on specific instructions proved invaluable for overcoming technical challenges.

All the prompts and code snippets are mentioned in a separate file named “**Prompt Engineering Documentation.**”

Final Thoughts:

The ranking system relies on multiple data, specifically, the candidates' final scores derived from a comprehensive analysis of their transcript data, emotional scores, and gaze patterns. This objectivity ensures a fair and unbiased assessment, eliminating the potential for subjective judgments or biases that might affect other selection methods.

Comprehensive Assessment: The final scores incorporate multiple facets of a candidate's qualifications. They reflect academic achievements, emotional intelligence, and engagement, providing a holistic view of a candidate's potential.

Using final scores as a ranking criterion demonstrates a commitment to data-driven decision-making. It aligns with modern recruitment practices, prioritising evidence-based strategies to identify the best-fit candidates.