Data Analytics Assignment

Kylie Diep - diep.soanh@gmail.com

1   Approach:

After exploring the data, data merging and cleaning are performed during data pre-processing in order to facilitate the further analysis. Data visualization and descriptive analytics are used for analysis. The analysis focuses on 5 key factors: CSAT, Suitability Recommendations, Plagiarism, Dropout, Candidates engagements.

1.1   Preprocessing:

1. **Linking and merging**: Source_B stores the candidate records of all status; therefore, the data set is merged with the other two by left-join.
   - Source_B left join Source_A on *prediction_id*, *cohortId*, *submittedAt*
   - Then left join CSAT on *prediction_id*
2. **Cleaning**:
   - Remove duplicates
   - Remove missing values for the following variables: *status*, *candidateId* for these records has no values for the analysis.
   - Remove *candidateID* = 0 since it is suspicious to have 3000 records in Source_B dataset.
   - Drop unnecessary variables: *gender*, *easl*, *cohortName*, *cohortId*, *profanityTest*, *plagiarismTest*, *jobId*, *recommend* (in source_B)

1.2   Analysis:

1. **CSAT**: Focus on *customer_satisfaction_score*, *comment_sentiment* and *polarity*
2. Suitability Recommendations:
   - Focus on how the *job_title*, personal traits, behavior traits reflect on *Recommend*. Proving the result given from FirstInterview is based on assumption of every job requires different personal and behavior traits beside the skill sets.
   - Prove the effectiveness of FirstInterview on identify the potential candidate and speed up the hiring process.
3. **Plagiarism**: focus on the proportion of candidates plagiarizing when participating in FirstInterview.
4. **Dropout**: focus on the dropout ratio, identify which recruiters and jobs have high dropout ratio.
5. **Candidates engagements**: focus on the submission ratio, identify which jobs have high submission ratio

# 2 Findings and Suggestions:

2.1 CSAT:

- High satisfaction scores (7 to 10) account for over 97% of the total feedbacks, indicating that the candidates have a good experience with FirstInterview.

DISTRIBUTION OF CUSTOMER SATISFACTION SCORE

- Positive sentiment dominates in the candidates' feedback which account for 88.38%. 10.36% of the feedbacks has neutral sentiment while negative feedbacks only accounts for 1.25% of the total feedback received from the candidates.

DISTRIBUTION OF COMMENT SENTIMENT
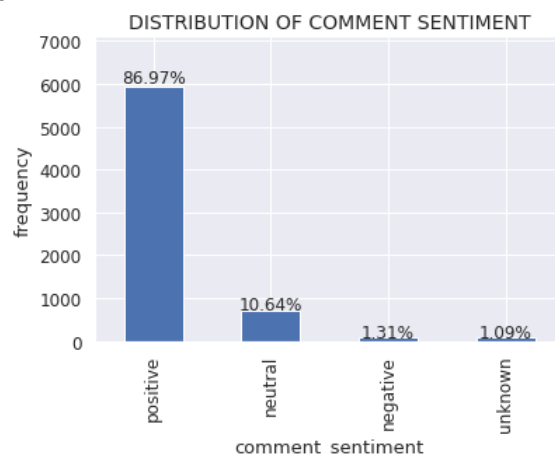
- Each sentiment has a wide range of polarity and satisfaction scores, which indicates that FirstInterview has some room for improvement.

Satisfaction score by comment sentiment

Polarity by comment sentiment
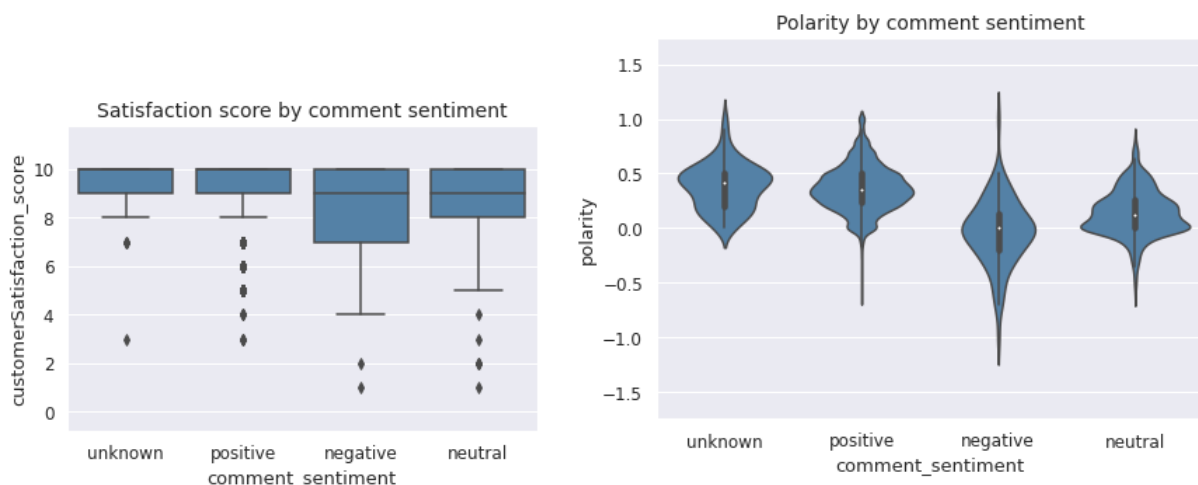
➯ **Future step**: Overall, the candidates have positive experience with FirstInterview. However, this result may not completely accurate for not all the users rate the product and provide feedback. To understand

why FirstInterview receives such a high positive feedback and how the product can be improved, it is necessary to futher investigate and analyse the detail of candidate comments.

⇨ **Suggestion**: Client feedback is also important for product development. It would be great to know whether the outcome of FirstInterview meet the business expectation and requirement.

2.2   Suitability Recommendations:

- FirstInterview eliminates nearly 40% of the total candidates, which helps the recruiter to reduce the number of unfit applicants and speed up the hiring process significantly.

DISTRIBUTION OF SUITABILITY RECOMMENDATION
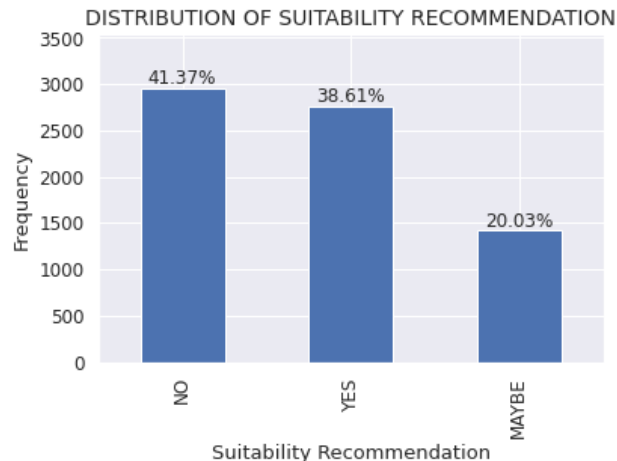
- The statistics summary and correlation plots (in Jupyter notebook) prove that the suitability recommendation given by PredictiveHive AI is unbiased. The recommendation is based on the assumption of every job requiring different personality and behaviour traits beside the skill sets.

⇨ **Future step:** Analyse the suitability recommendation based on skill sets and other factors.

⇨ **Suggestion:** It would be great if there is data about successful candidates in FirstInterview got employed by the business so that we can show to the clients how helpful and successful FirstInterview is in hiring process beside speeding up the hiring process by reducing the number of applicants.

2.3   Plagiarism:

- Majority of the candidates do not plagiarise, reflecting nearly 82%. There is 16.26% undetermined while only 2% of candidates are identified as plagiarism.
- There is an interesting insight about plagiarism. Most of the candidates who have significantly high value in their personality traits are undetermined in plagiarism. The visualization can be found in Jupyter notebook.


DISTRIBUTION OF PLAGIARISM TYPE
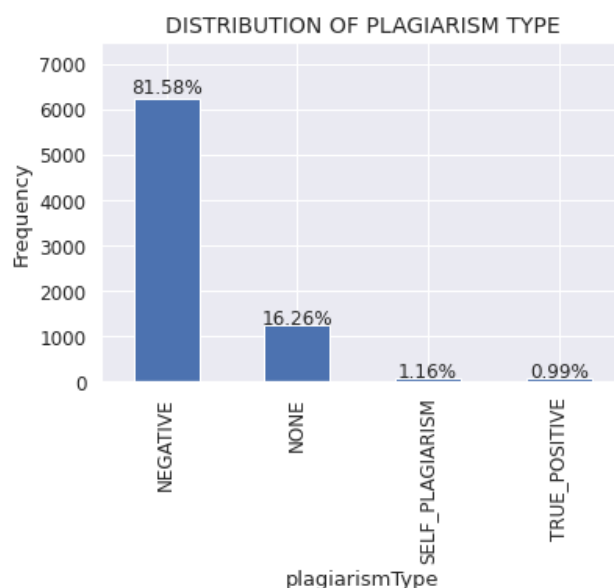
2.4 Dropout:

- The proportion of dropout when candidates participated in FirstInterview is only 2.3% while the completion rate is 97.7%.
- There is a certain number of recruiters and jobs having more dropouts compared to others.
- ⇨ **Suggestion:** The business needs to understand why the candidates dropped out by looking into their feedback.
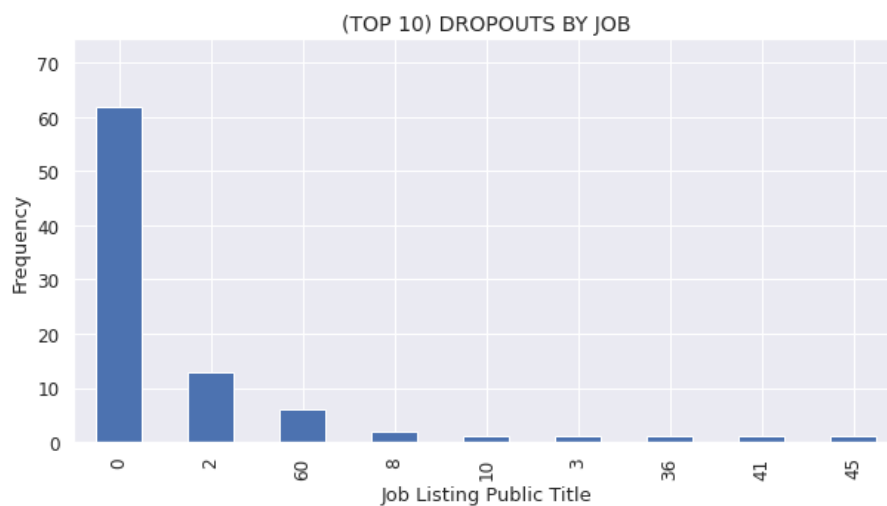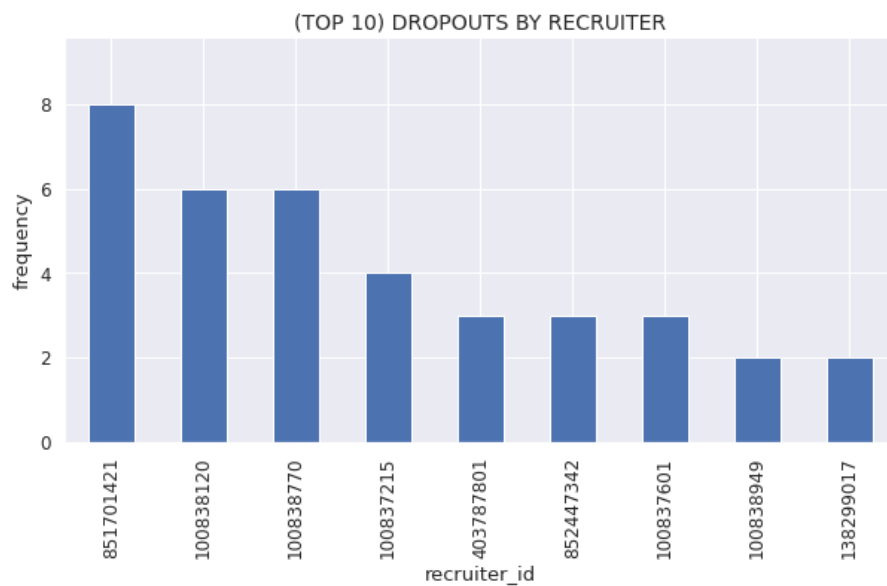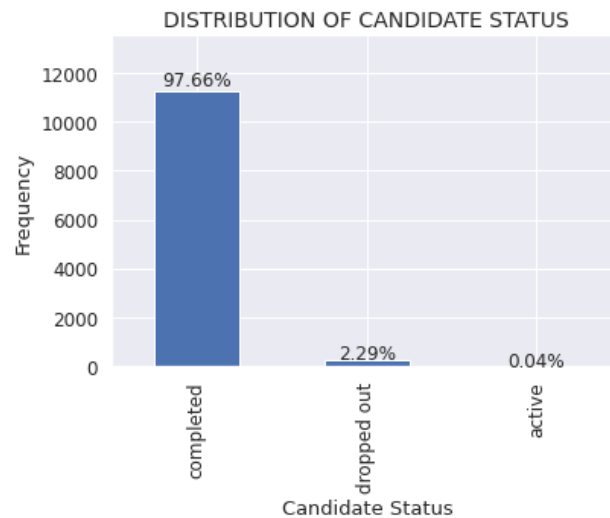
2.5   Candidates engagements:

- People are likely to apply for job in April and end of August while the number of job available increase around mid May and September. It is true especially for graduates who are desperate to look for employment after their study. However, only a certain number of jobs attract candidates to apply and not all the candidates completes their first round of interview.
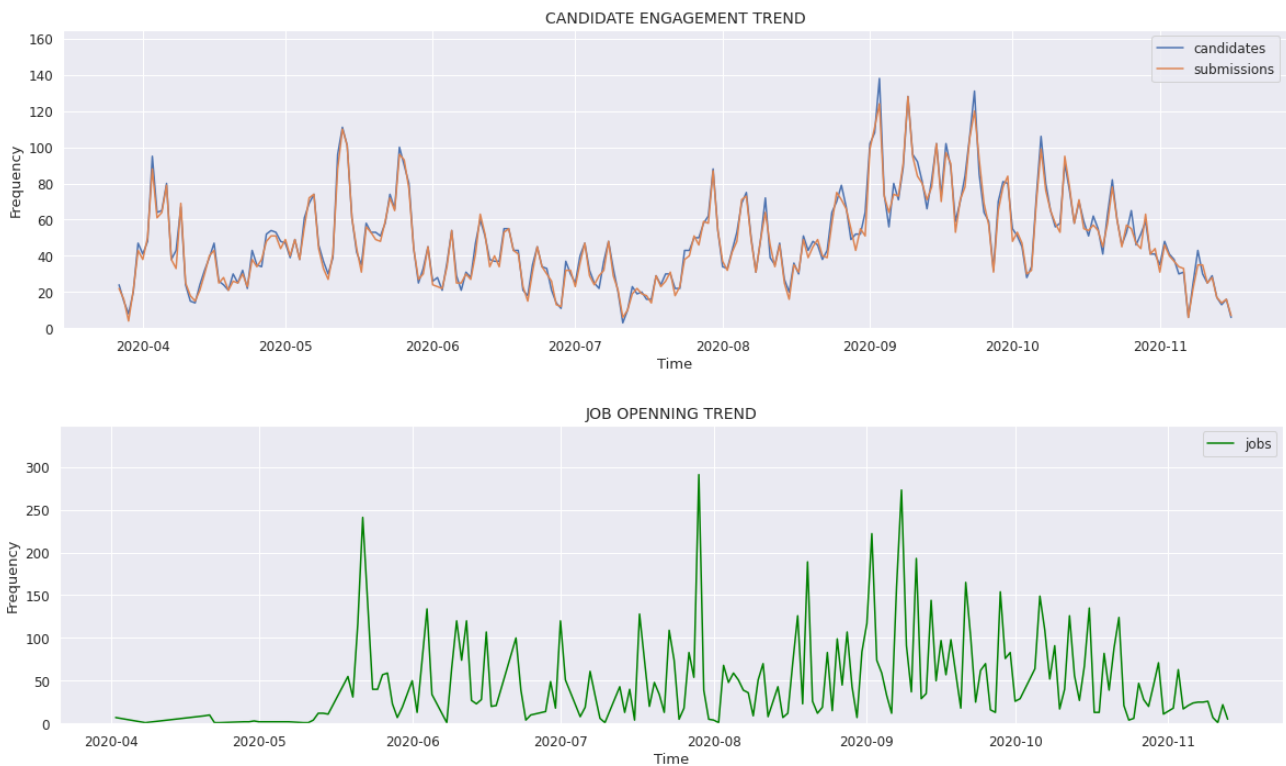
⇨ **Suggestion:** Deep dive in user feedback to make improvement in product development to enhance the candidate engagements.