

GRADSG Survey Respondents Analysis Report

Introduction

Survey data plays an important role in informing organisational decision-making, particularly in understanding perceptions, motivations, and behavioural patterns of respondents. However, large-scale surveys often result in complex datasets that include structured responses (MCQ, 1-10 ratings, etc.) and unstructured inputs (open-ended text), which are not straightforward to analyze. In addition, poorly designed survey questions or redundant questions may reduce response quality, leading to incomplete submissions and less useful insights.

This study aims to address what patterns exist in respondent behaviour, which factors influence employer attractiveness, where do respondents disengage or fail to complete surveys, and provide meaningful insights and suggestions for further surveys.

Dataset Overview

The analysis is conducted using the GRADSG dataset, which captures undergraduate and graduate students' perceptions of employer attractiveness, career motivations, and informational needs for potential employers. The dataset includes a few categories:

"Response ID", "Time Started", "Date Submitted", and "Status" columns help indicate the time taken and survey completion.

"SessionID", "User Agent", "Tags", "Language", "Country", "City", "Latitude", "Longitude" columns provide contextual information for behavioural pattern differences across respondent groups.

Questions regarding institution of study, year of study, highest qualification upon graduation, field of study, nationality, and gender enable segmentation analysis to identify academic clusters and their perceptions of employers.

There are also "pick 3" and open-ended questions relating to the employer, i.e. respondents' perception of the employer, what they wish to learn more about the employer, the attractiveness of the employer, and respondent's motivation to apply to this employer.

Methodology

Upon inspecting the dataset, we dropped "SessionID", "User Agent", "Tags", "Language", "City", "Latitude", "Longitude" columns as it is quite redundant to use in our analysis. We had also renamed most of the question columns to make it concise. Luckily there weren't any duplicate rows, but we kept the missing values of some columns as it helps with the incomplete survey analysis.

Following data preparation, we proceeded with Exploratory Data Analysis (EDA) and feature engineering:

a) Survey status and time taken analysis:

We calculated the proportions of 'Complete', 'Partial', and 'Disqualified' surveys to understand the overall completion rates. We then analyzed the `time_taken` for each survey status, calculating minimum, maximum, and average times, and visualizing the distributions using box plots. This allowed us to observe insights such as the prevalence of short disqualification times and the existence of extremely long completion times for some surveys.

b) Association analysis:

To understand relationships between variables and survey status, we generated cross-tabulations and bar plots to visualize the distribution of 'Complete', 'Partial', and 'Disqualified' statuses across key categorical variables like `country`, `year_of_Study`, `highest_qualification`, `major`, `gender`, and `perception`. We computed Mutual Information scores to quantify the dependency between all relevant features and the status column, identifying the `scale(1-10)` attractiveness rating and `time_taken` as the strongest predictors of survey status. For numerical features, a Kruskal-Wallis test was conducted to assess significant differences in the `scale(1-10)` rating across different survey statuses.

c) Perception and attractiveness rating analysis:

We specifically investigated the relationship between respondents' initial perception of the organization and their `scale(1-10)` attractiveness rating. Descriptive statistics (mean, median, standard deviation) were calculated for `scale(1-10)` grouped by perception categories. A violin plot was used to visualize these distributions, and a Kruskal-Wallis H-test confirmed a statistically significant difference in ratings based on perception.

d) Natural Language Processing (NLP) for question analysis:

To identify redundancies and thematic structures within the survey questions, we applied NLP techniques:

- Survey questions were transformed into numerical embeddings using a SentenceTransformer model.
- A cosine similarity matrix was generated from these embeddings to quantify semantic similarity between all pairs of questions. Based on a threshold of 0.7, redundant or highly similar question pairs were identified.
- K-Means clustering was applied to the question embeddings to group questions into thematic categories (e.g., General Demography, Applicant's Expectation on Employer, Education Status, Employer Attractiveness).

e) Demographic and behavioural segmentation analysis:

- Demographic segmentation:
 - Demographic features such as 'university', 'year_of_Study', 'highest_qualification', 'major', 'nationality', and 'gender' were extracted from the `grad_clean` dataset.

- These categorical demographic features were then one-hot encoded and missing values were filled with 'Missing'. The encoded data was subsequently scaled using StandardScaler to ensure all features contributed equally to the clustering process.
- The Elbow Method (WCSS) and Silhouette Score were employed to suggest an optimal number of clusters. While the Silhouette Scores were relatively low (indicating less distinct clusters), a choice of k=3 was made to target broad audience segments for the renewed survey, rather than highly niche ones.
- Behavioural segmentation:
 - Behavioural features were selected, including 'perception', 'types_of_roles', 'career_progression', 'compensation', 'worklife_balance', 'interview_process', 'scale(1-10)', and 'motivation_factor'.
 - Similar to demographic data, these features were one-hot encoded, with missing values filled, and then scaled using StandardScaler.
 - K-Means clustering was applied with k=4 to identify distinct behavioural segments.

Results

- a) Our analysis of complete surveys revealed a wide range in completion times. While the average time taken to complete a survey is approximately **4 hours and 2 minutes**, we observed extreme outliers, with some surveys taking as long as **41 days, 13 hours, and 18 minutes**. This indicates that many respondents are likely not completing the survey in a single session, possibly leaving it open for extended periods. Such prolonged durations can affect the quality of responses, as respondents might lose focus or their perceptions could change over time.

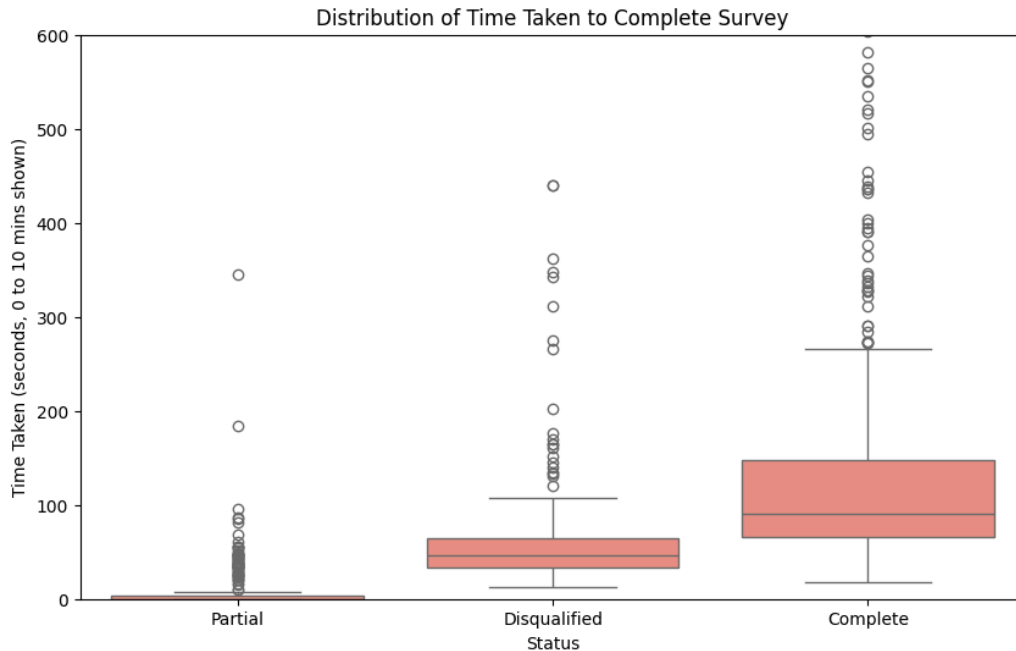


Figure 1

- b) There is a statistically significant difference in the mean attractiveness ratings across different categories of “perception” towards the organization. A Kruskal-Wallis H-test confirmed a significant difference (H-statistic = 280.116, p-value = 0.0000). Respondents who were already familiar with and would consider the organization as a potential employer (“I’m familiar with the organisation...”) gave the highest mean rating of 7.37. Those with only name recognition (“I recognise the organisation by name...”) gave a mean rating of 5.88. Those not familiar enough to form an opinion (“I’m not familiar enough with the organisation...”) gave the lowest mean rating of 5.16. **Prior familiarity and a positive initial impression are strongly correlated with higher attractiveness ratings.** Marketing and employer branding efforts could focus on increasing awareness and fostering positive perceptions before survey distribution.

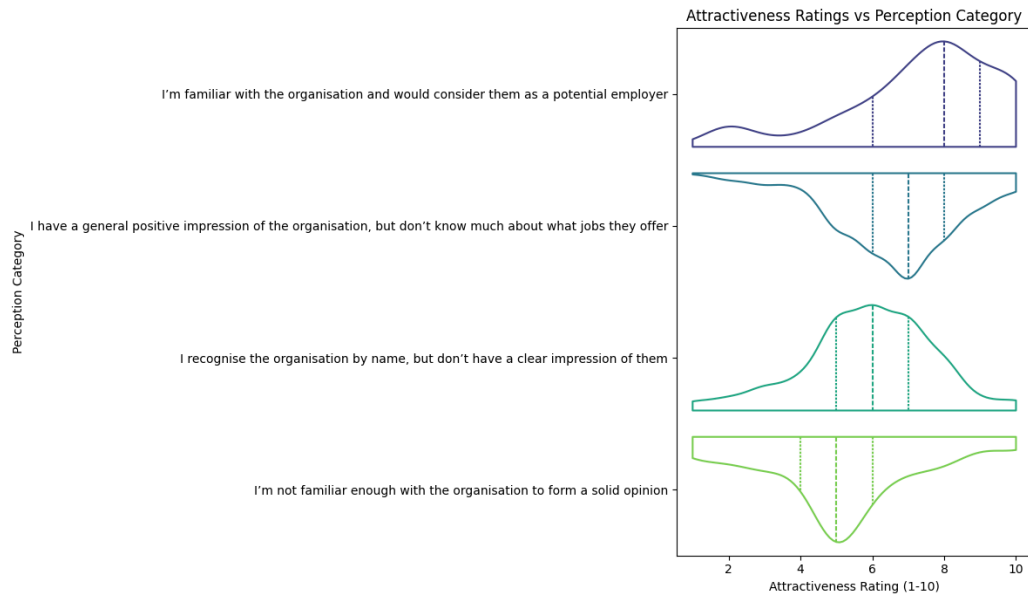


Figure 2

- c) NLP-based similarity analysis revealed multiple question pairs with cosine similarity scores exceeding **0.70**, which indicates substantial **conceptual duplication**. In particular, questions within the “What do you wish to learn more about...” pick-3 section (covering `types_of_roles`, `career_progression`, `compensation`, `worklife_balance`, `interview_process`, `other_1`, and `other_2`) exhibited inter-question similarity scores ranging from 0.72 to 0.82. Additionally, the `motivation_factor` question and its corresponding open-ended “other” response showed a notably high similarity score of 0.85, which further reinforces the presence of overlapping intent.
- d) The survey questions naturally cluster into four distinct thematic areas. K-Means clustering on question embeddings identified **General Demographics** (nationality, gender), **Applicants’ Expectations of Employers** (perception, `types_of_roles`, `career_progression`, `compensation`, `worklife_balance`, `interview_process`, `other_1`, `other_2`), **Educational Background of Applicants** (`high_ed`, `year_of_study`, `highest_qualification`, `main_subject`), and **Employer Attractiveness** (scale (1–10), `motivation_factor`, other). These groupings reflect clear conceptual boundaries across question types, indicating a coherent underlying structure in the survey design.
- e) The Mutual Information analysis indicates that the **scale (1–10) employer attractiveness rating** and the **time taken to complete the survey** are the strongest predictors of survey status (Complete, Partial, or Disqualified). The scale (1–10) variable exhibits a very high Mutual Information score of **0.606**, largely because it is typically

answered only in fully completed surveys, making its presence a strong signal of completion. Similarly, time_taken shows a substantial score of **0.353**, which shows respondent engagement duration as a key indicator of whether a survey is completed or not. In contrast, most demographic variables display considerably lower Mutual Information scores, suggesting they play a much smaller role in determining survey status. Overall, these findings highlight that **engagement levels and the completion of core evaluative questions are far more influential drivers of survey completion than respondent demographics.**

- f) The K-Means clustering analysis with $k=3$ successfully identified three demographically distinct segments among the respondents. **Cluster 0**, representing approximately **9%** of the sample ($N=240$), consists primarily of students from **Yale-NUS College**, though this group notably lacks detailed demographic information across several key variables including year of study, highest qualification, major, nationality, and gender. This data gap suggests either a unique profile for these respondents or incomplete data capture for this particular institution. **Cluster 1** accounts for roughly **30%** of respondents ($N=797$) and is characterized by a diverse university representation, predominantly from **Nanyang Technological University (NTU)**, with smaller proportions from **Singapore Management University (SMU)** and **Yale-NUS College**. This segment consists mainly of undergraduate students pursuing Bachelor's degrees across various years of study, with their academic interests concentrated in Business/Management, Engineering, Economics, and IT & Technology. The demographic profile skews toward **Singaporean citizens or permanent residents**, with a slightly higher proportion of **female** respondents. **Cluster 2** forms the largest segment at approximately **60%** of the total sample ($N=1577$) and is almost exclusively composed of **National University of Singapore (NUS)** students (99.4%). This cluster exhibits a distinct academic focus on Natural Sciences, Mathematical Science/Statistics, and Medicine/Dentistry, differentiating it from the more business and technology-oriented Cluster 1. Similar to Cluster 1, this segment is predominantly Singaporean/Singapore PR with a higher proportion of female respondents, suggesting certain demographic consistencies across the majority of the surveyed population despite differences in institutional affiliation and academic focus.

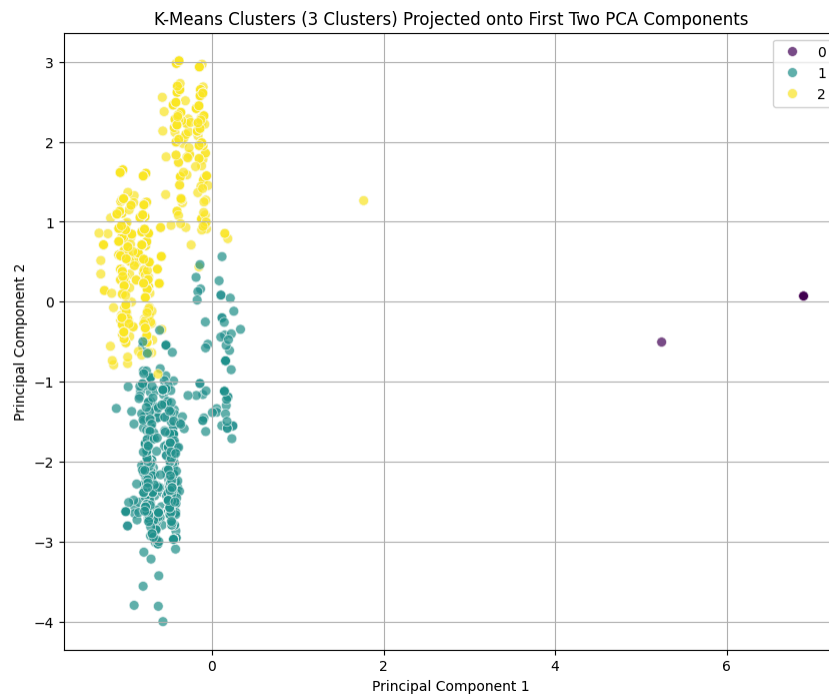


Figure 3

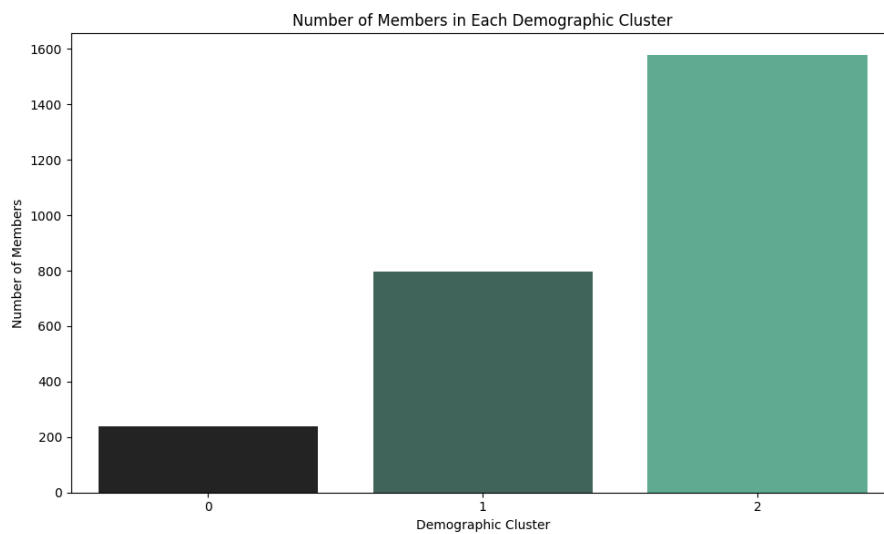


Figure 4

- g) The behavioral segmentation revealed **four** distinct clusters with varying levels of organizational engagement. **Cluster 0**, the largest group with **945** respondents, represents **moderately aware individuals** who assign attractiveness ratings of 6-7 and are motivated by meaningful work impact, career growth, and work-life balance. This segment presents a prime opportunity for targeted messaging emphasizing the

organization's mission and development programs. **Cluster 1**, containing **766** respondents, is characterized by **extensive missing data**, suggesting incomplete survey responses that offer minimal behavioral insights and indicate potential issues with survey design or completion rates. **Cluster 2** comprises **354 unfamiliar or skeptical respondents** who consistently rate the organization at 5.0 despite sharing similar motivators with Cluster 0. Their low familiarity combined with moderate ratings suggests barriers to engagement that require foundational brand awareness campaigns and clear value proposition communication. **Cluster 3** represents **549** highly engaged individuals who **demonstrate strong familiarity with the organization** and consistently provide the highest attractiveness ratings of 8-10. Motivated by meaningful work impact, career growth, and job security, this segment represents the organization's strongest advocates and most promising recruitment pool, warranting strategies that leverage their positive sentiment and align opportunities with their key motivators.

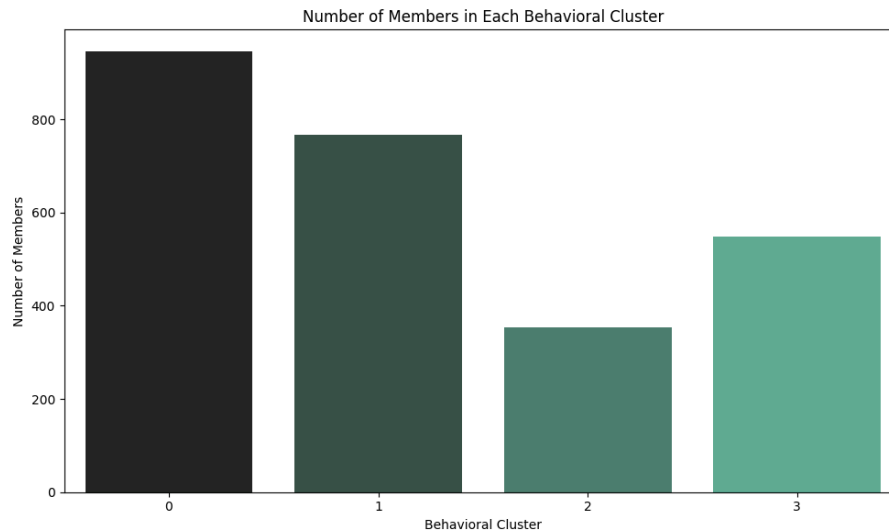


Figure 5

Insights

- To ensure higher data quality and encourage more focused responses, consider implementing a **session timer** for the survey. This timer could be set based on the average completion time, with a reasonable buffer (e.g., 6-8 hours or a full day depending on the survey complexity), prompting users to complete the survey or saving their progress within a defined period.

- b) People who already know the organization and can imagine themselves working there consistently rate it more favorably, while limited or no familiarity translates into lower scores. This suggests that the survey is not just measuring employer appeal in isolation, but also the effectiveness of pre-existing awareness and brand perception. Strengthening employer branding and increasing meaningful exposure before engaging respondents would likely raise baseline attractiveness ratings and lead to more informed feedback.
- c) Respondents are likely interpreting several questions as asking essentially the same thing, especially within the *“What do you wish to learn more about...”* section and between the motivation_factor question and its open-ended “other” counterpart. This suggests that the survey is capturing overlapping signals rather than distinct areas of interest which can dilute the quality of responses and contribute to unnecessary survey length. Streamlining these items reduce respondent cognitive load and help ensure that each question elicits a more clearly differentiated insight which would make the resulting data more actionable and easier to interpret.
- d) The survey already aligns closely with how respondents mentally organize information about employers, even without explicit sectioning. However, by formally grouping questions into these four themes, the survey can better match respondents’ cognitive flow, making it easier to answer consistently and thoughtfully. At the same time, this thematic structure would allow analysts to isolate and interpret drivers of employer attractiveness more clearly which may improve both respondent experience and analytical clarity.
- e) Survey completion is primarily driven by respondent engagement rather than who the respondents are. The presence of the core attractiveness rating and the amount of time spent on the survey are the strongest indicators of whether a survey is completed, while demographic characteristics have little influence on survey status. This shows that completion behavior is linked more to how respondents interact with the survey than to their background. Streamlining questions, improving clarity, and making the survey feel purposeful from the start are likely to increase completion rates more effectively than targeting specific demographic groups.
- f) The demographic segmentation analysis reveals some improvements for future surveys. First, Cluster 0’s significant missing data, primarily from Yale-NUS students, indicates a need for mandatory demographic fields or more inclusive response options that accommodate diverse educational models like liberal arts programs. Reducing these data gaps would enable more comprehensive analysis of all respondent segments. The distinct institutional profiles of Clusters 1 and 2 (dominated by NTU and NUS) suggest opportunities for targeted outreach strategies. NTU students concentrate in Business, Engineering, and IT, while NUS students focus on Natural Sciences, Mathematics, and Medicine. Future surveys and recruitment efforts could leverage these differences

through tailored messaging that resonates with each institution's academic strengths and student interests. Major-specific concentrations also likely indicate varying priorities regarding employer attractiveness. Future surveys should incorporate conditional logic or discipline-specific questions to explore how factors like compensation, work-life balance, and career growth vary across academic fields, enabling employers to develop more refined engagement strategies. Finally, the predominance of Singaporean/PR and female respondents in Clusters 1 and 2 suggests underrepresentation of international students and other demographic groups. If broader representation is desired, survey distribution should expand through diverse student organizations, international networks, and potentially multilingual formats to capture a more comprehensive view of the student population.

- g) The behavioral segmentation suggests tailored engagement strategies for each cluster. Clusters 0 and 2 require targeted messaging that builds deeper organizational understanding and showcases unique value propositions around impact, career development, and work-life balance, while Cluster 3 benefits from reinforcement of existing positive perceptions and highlighting opportunities aligned with their motivations. The presence of Clusters 0 and 2 indicates a need for enhanced brand awareness initiatives to improve recognition and clarify the organization's employer value proposition to broader audiences. The substantial size of Cluster 1 points to potential survey design issues which suggests that reviewing question clarity, survey length, and completion flow could reduce missing data in future iterations. Cluster 3 may also represent a valuable asset. Analyzing what attracts these highly engaged individuals can inform broader employer branding and recruitment strategies to attract similar high-quality candidates.

Conclusion

This analysis demonstrates how exploratory data analysis and natural language processing can be used to turn complex survey data into actionable insights and identify opportunities to improve survey design. Our findings show that survey completion is driven primarily by respondent engagement and survey structure rather than respondent demographics, highlighting the importance of clear, purposeful, and well-paced questions. Prior familiarity strongly correlates with respondent perception of employer attractiveness ratings, suggesting that survey responses are not influenced only by the questions themselves but also by respondents' existing perceptions. NLP-based similarity and clustering analyses revealed areas of conceptual redundancy and clear thematic groupings within the survey questions, offering concrete recommendations for streamlining question design and improving respondent experience. Overall, these findings emphasise that effective survey design, clear question differentiation, and thoughtful structuring are critical to improving data quality and interpretability. The insights from this analysis provide a foundation for building internal analytics tools that

support better survey design, more meaningful insights, and more informed organisational decision-making.