NUS SDS Datathon 2026

TEAM NAME

John **A0322572J**
Declan **A0323192L**
Sean **A0323255L**
Richie **A0323259A**

Abstract

Surveys are vital for gathering feedback, but extracting reliable insights from high-volume responses remains challenging. Traditional manual analysis often overlooks nuanced respondent behaviour and fails to identify survey design weaknesses.

We developed an end-to-end survey analytics framework to help research teams transform raw responses into actionable intelligence. Processing 2,632 survey responses, we cleaned invalid entries, standardised demographics, and applied exploratory analysis to uncover patterns and correlations in survey structure. XGBoost classification achieved 0.807 ROC-AUC in predicting high-attractiveness responses (ratings ≥8), capturing subtle interactions between demographic and perceptual variables.

An interactive dashboard built with React and AI-powered insights (via LLM integration) enables surveyors to visualise results, identify drop-off points, and receive automated recommendations for survey improvement. The tool addresses key operational needs: detecting question redundancies, understanding incomplete responses, segmenting respondents by behaviour, and synthesising employer attractiveness drivers from the applicant perspective.

This framework bridges the gap between raw survey data and strategic decision-making, providing surveyors with practical, interpretable intelligence for continuous survey optimisation.

Introduction

Surveys are vital to organisations in gathering feedback and insights. It drives decision-making through the data it provides. With the sheer volume of survey data collected by organisations, ensuring that data is accurate, and thereafter deriving purposeful and actionable insights remains a noteworthy challenge.

Traditionally, survey data analysis is conducted manually through basic interpretation or statistical summaries. This tends to overlook many factors such as varying respondent behaviour. This can differ in many ways. Where the respondent is from, the respondent's age and even how the questions in the survey are set—all variables that affect how people respond to surveys.

To address these issues, our group intends to leverage modern data analysis techniques to organise raw survey data. Our solution strives to not only curate a meaningful and reliable dataset, but also identify correlations and redundancies in survey questions through hidden nuances. Ultimately, this will close the gap between raw data and strategic decision making.

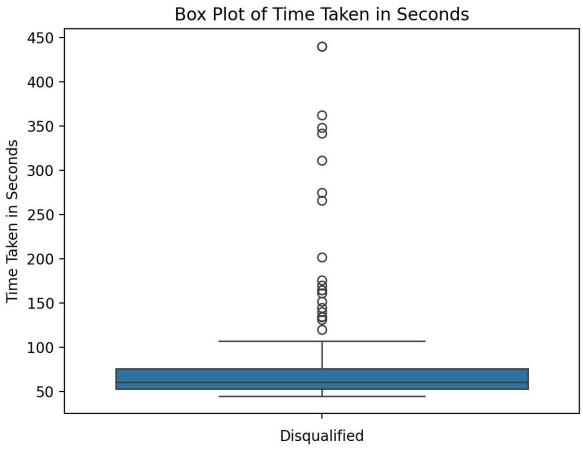
The raw survey dataset we are using in this study consists of 2632 responses. We intend to consider Age, Nationality, School & Faculty, Gender and Survey Duration in the data to normalise the survey results.

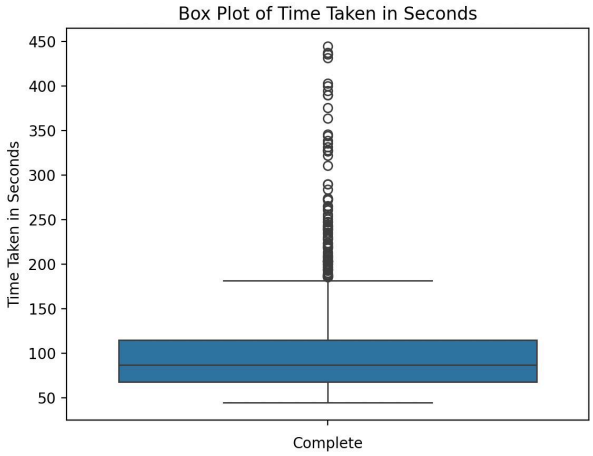
Exploratory Data Analysis

Data Filtering Methodology

The raw GRADSG dataset was firstly imported, then standardised to clean column names. Thereafter, a new variable, time_taken_seconds, was derived from the given start and end times in the dataset to measure how long each respondent took to complete the survey.

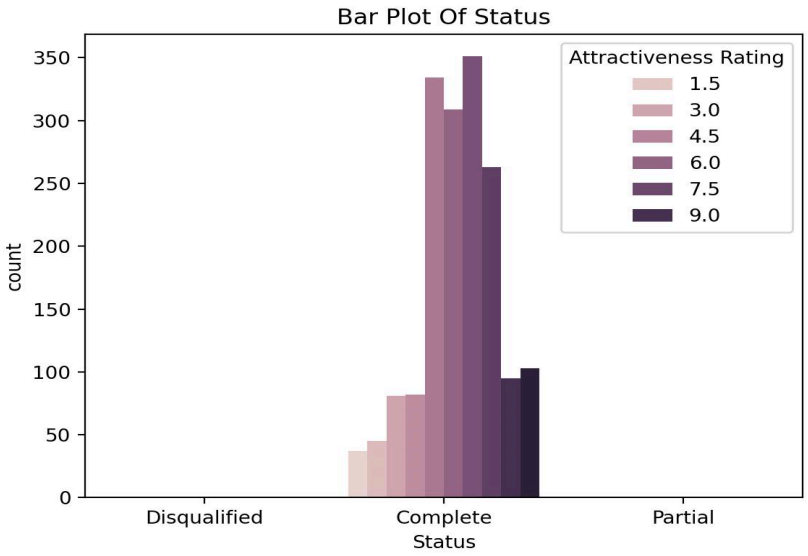
Based on the survey questions in the dataset, below 45 seconds is an unrealistic response time, reflecting that the respondent did not have meaningful participation in the survey. Hence, the responses below 45 seconds were excluded together with responses that did not answer any questions. The respective distributions of responses that were disqualified and that were accepted are shown below.





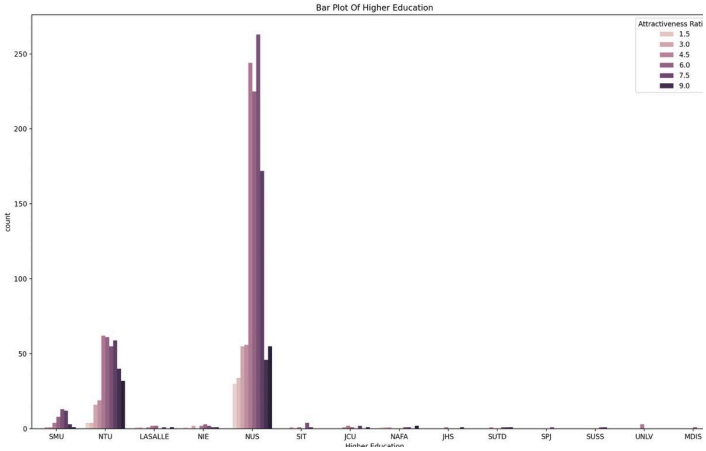
Data Visualisation

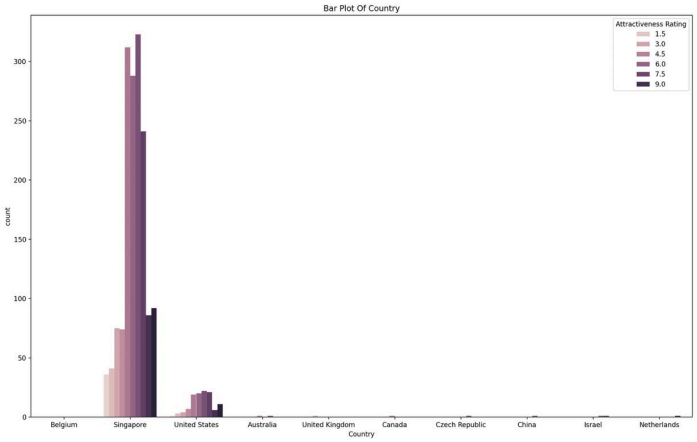
The distribution of the 1976 responses from the cleaned dataset is shown below.



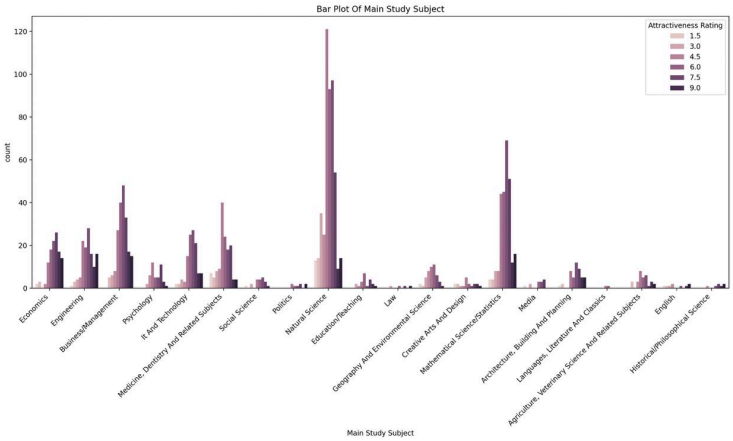
Demographic Data

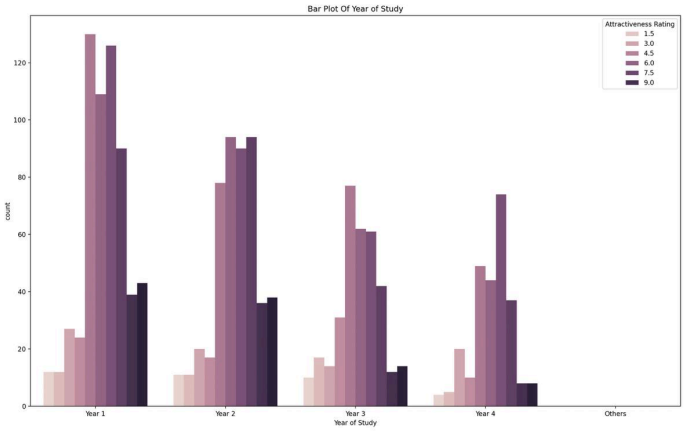
In order to further analyse the correlation between different variables and uncover nuances in the data, respondents' answers are grouped according to their answered Attractiveness Rating, and each individual variable below.



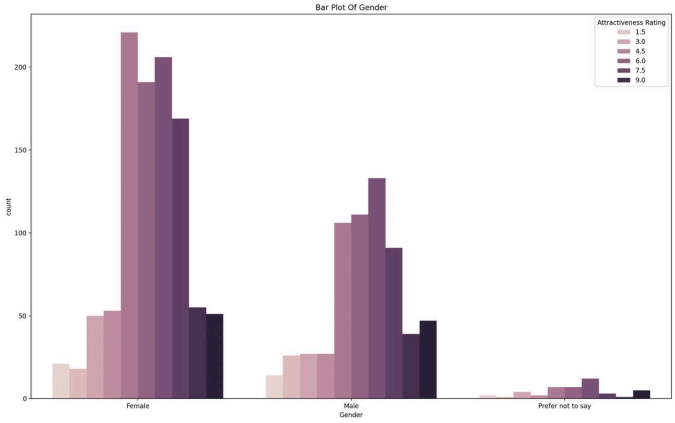


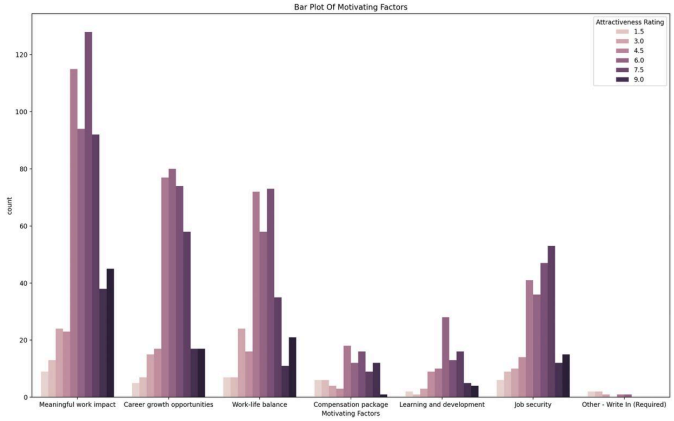
Based on the bar plots for the respondents' institution and country, the National University of Singapore (NUS), and Singapore respectively make up the majority of the participants. Most of the remaining participants are from either Nanyang Technological University (NTU) or Singapore Management University (SMU).





Within their respective institutions, the majority of the participants are studying Natural Science. There are more year 1 students participating compared to other years.





Based on the bar plots above, meaningful work impact is most often selected as a motivating factor. And lastly, the gender ratio of respondents is skewed towards females.

Demographic Summary

In conclusion, most survey respondents are females from Singapore and study at either NUS, SMU or NTU with the most common major being Natural Science. There is a larger proportion of Year 1 students. Given this demographic concentration, we can expect the data to be skewed towards the majority demographic.

Model Selection

The following models were considered:

- Logistic Regression
- Random Forest
- XGBoost
- Naive Bayes
- K-Nearest Neighbours

- Linear Regression

Tuning Methodology

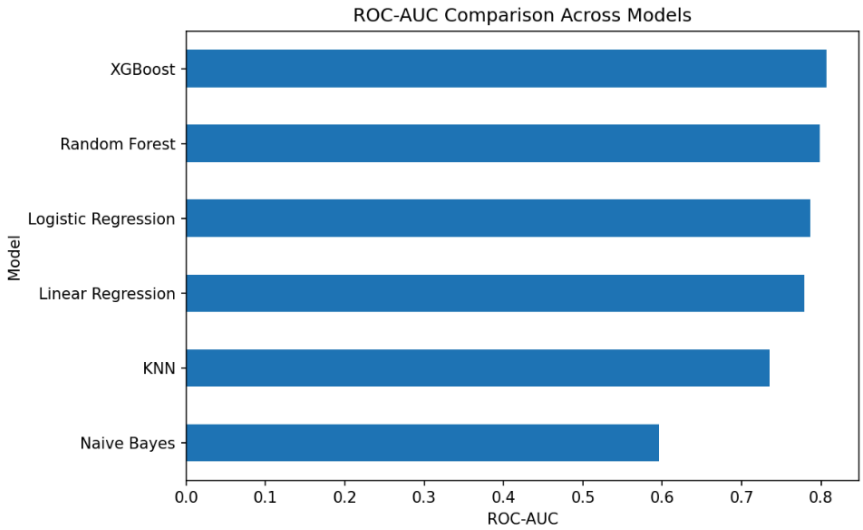
The predictive models were trained through a unified preprocessing and tuning pipeline with the cleaned dataset. Numerical variables were imputed using median values and standardised whilst categorical variables were imputed using the most frequent category and one-hot encoded.

The dataset was partitioned using an 80-20 stratified split on the attractiveness variable to preserve class proportions. Model hyperparameters were tuned using 5-fold cross-validation, with ROC-AUC prioritised.

ROC-AUC was prioritised as survey outcomes tend to be nuanced and the distinction between responses is gradual rather than separated. Thus, it is paramount that the model consistently assigns higher predicted scores to genuinely high-attractiveness instances relative to lower ones. Survey response distributions may also exhibit class imbalance, resulting in misleading results from other metrics such as F1-Score. Hence, when ROC-AUC is prioritised, models will be judged according to their separation capability; ability to distinguish subtle variations in respondent perception.

Results & Model Comparison

Model	ROC-AUC
XGBoost	0.807
Random Forest	0.798
Logistic Regression	0.786
Linear Regression	0.779
KNN	0.730
Naive Bayes	0.596



Naive Bayes - 0.596

Naive Bayes assumes that features are conditionally independent given the class which should be too strong for a survey dataset. Hence, it is expected that Naive Bayes produced the weakest results with 0.596; only slightly better than random selection.

K-Nearest Neighbours - 0.730

KNN classifies new data points based on the k-nearest, labelled neighbours. In this instance where categorical variables are one-hot encoded, distance becomes less informative, hence it does less well compared to tree ensembles.

Linear Regression - 0.779

Linear regression can evaluate whether each individual score ranks positives over negatives if its outputs are treated like a continuous score. The ROC-AUC of 0.779 shows that there is some linear signal.

Logistic Regression - 0.786

Also a linear model like Linear Regression, it sends an obvious signal when variables are correlated but does not capture more complex interactions. Hence, it would not be the best model to capture smaller nuances.

Random Forest - 0.798

Random Forest is a tree ensemble with randomness included through feature subsampling. It is thus more suitable to detect irregularities and non-linear differences without feature engineering.

XGBoost - 0.807

Compared to Random Forest, XGBoost builds trees sequentially and learns from previous model's mistakes. It likely has a high ROC-AUC rating because it separates regions better, resulting in subtle interactions being made more apparent. This shows that XGBoost, incrementally learning small, targeted rules, is the most suitable model compared to that of independent trees or using a linear boundary.

Model Selection Limitations

Model Limitation

While XGBoost achieved strong performance (0.807 ROC-AUC), its black-box nature limits interpretability. Feature importance provides general patterns but individual predictions remain difficult to explain to stakeholders.

Sample Representativeness

The dataset exhibits demographic imbalances including gender (67% female, 33% male) and institutional concentration (85% from NUS/SMU/NTU). This may limit generalizability to the broader student population and could bias model predictions toward majority demographic patterns.

Data Quality & Response Patterns

The 45-second minimum response time threshold, while removing clearly invalid responses, remains somewhat arbitrary. Survey fatigue is evident with a 35% drop-off at Question 7, suggesting later questions may have lower response quality.

Measurement Validity

All data is self-reported, introducing potential social desirability bias. Students' stated preferences for "meaningful work" may differ from actual career decisions. The single-item 1-10 attractiveness scale may not fully capture the multidimensional nature of employer attractiveness.

Dashboard

The interactive Survey Analytics Intelligence dashboard designed to be used by non-technical stakeholders. It makes sense of survey results, extracts insights and obtains data-driven recommendations, in legible terms for the average individual.

It was built using React with a Node.js 18.x runtime. The cleaned dataset can be embedded into the frontend, ensuring smooth usage. Using the Groq API, running the Llama-3.1-8b-instant model, the user can understand survey insights through an AI-assisted insights module that automates interpretation and provides recommendations based on patterns present in the survey dataset.

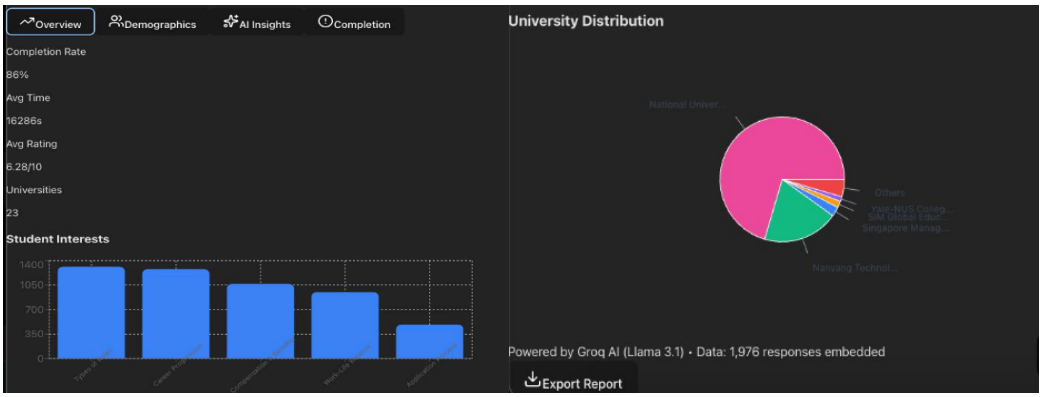
The interface consists of four pages, namely:

- Overview
- Demographics
- AI Insights
- Completion Analysis

Overview

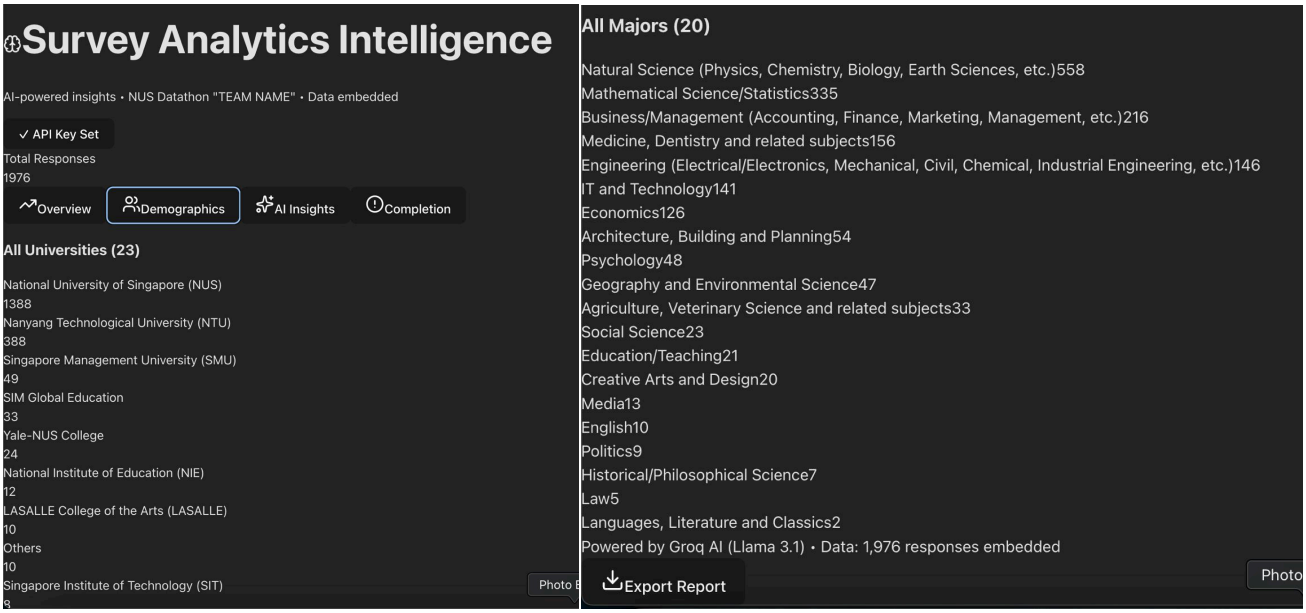
The overview page shows a high-level summary of survey results through interactive bar & pie charts as seen below. This allows the user to easily grasp key patterns with summary statistics such as response numbers, average attractiveness ratings and response distribution, and thereafter choose which aspects they would like to focus on.

Users can also export processed data and visual outputs to facilitate offline review, reporting or further analysis.



Demographics

The Demographics page provides the full breakdown of respondent background characteristics including that of respondents' schools, faculties and majors. These visualisations allow the user to quickly identify dominant respondent groups and demographic skew. This enables a more informed interpretation of the results and supports more targeted outreach efforts for future campaigns.

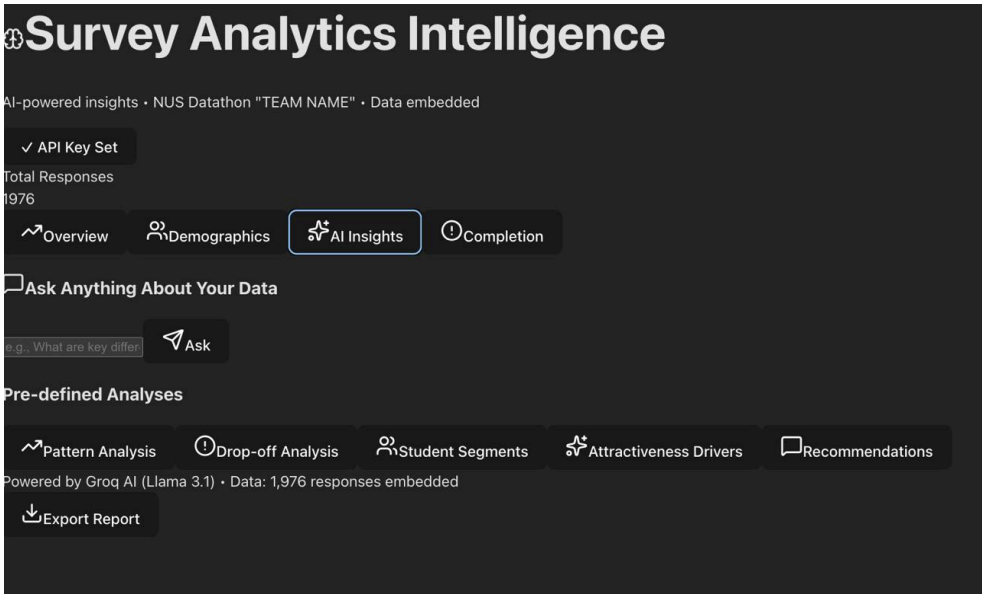


AI Insights

The AI insights page provides actionable suggestions through AI-powered analysis. This page includes:

- Pre-defined prompts for common analytical queries
- A text box for custom queries by the user

The integrated language model analyses survey patterns and generates insights and suggestions, all in plain language. Such patterns can include question phrasing improvements or identifying engagement issues. This unlocks guided analytical support without manually interpreting charts of statistical outputs, perfect for non-technical users.

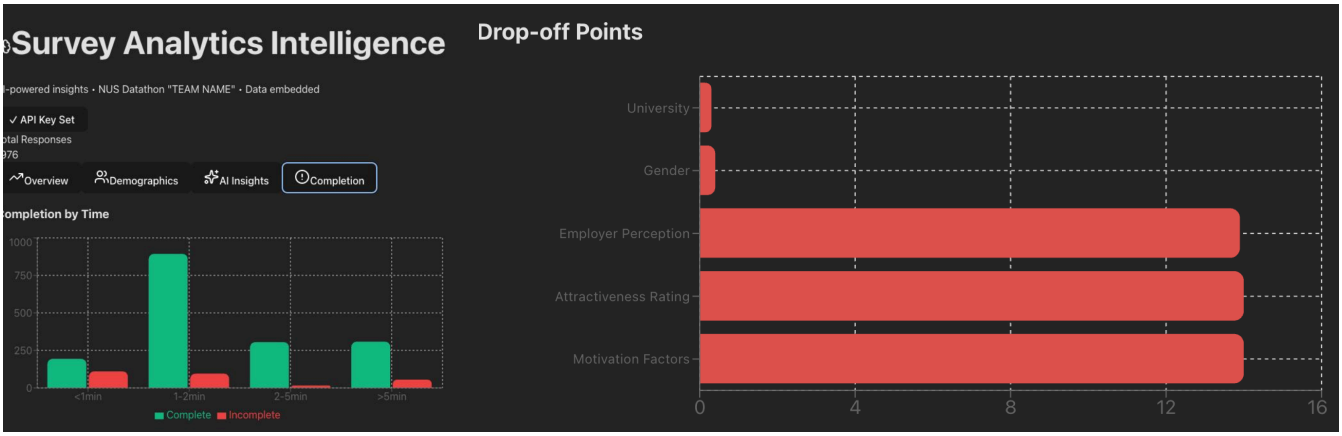


Completion Analysis

The completion page dives deeper into respondent behaviour and survey quality metrics. It visualises the expected survey duration, with zero-second entries removed to exclude invalid submissions.

The main highlight of this page is the drop-off rate analysis by question section. This section highlights where respondents disengage most frequently. This allows the user to identify unclear

or overly-long question segments and refine survey structure in order to improve completion rates in the future.



Conclusion

This project successfully delivered an end-to-end survey analytics framework that transforms raw survey data into strategic intelligence for research teams and surveyors. Through systematic data cleaning, exploratory analysis, and demographic profiling, we identified key respondent patterns, sampling biases, and variable correlations that traditional methods often miss.

Our predictive modelling approach, guided by ROC-AUC optimisation, achieved 0.807 with XGBoost, demonstrating superior capability in capturing nuanced, non-linear interactions between survey variables. This represents a significant improvement over conventional statistical summaries, enabling more precise identification of high-attractiveness response patterns.

The Survey Analytics Intelligence dashboard translates these complex analytical outputs into practical value. By combining interactive visualisations, AI-assisted interpretation, and completion analysis, the tool empowers non-technical stakeholders to extract actionable insights, diagnose survey design weaknesses, and optimise future data collection efforts, all without manual statistical expertise.

Beyond technical performance, this framework addresses real operational challenges: reducing analysis time from days to minutes, surfacing hidden patterns in respondent behaviour, and providing evidence-based recommendations for survey improvement. The 35% drop-off identified at Question 7, for instance, directly informed redesign priorities.

Future enhancements could expand impact further by integrating natural language processing for open-ended responses, enabling multi-survey comparative analysis, and implementing real-time validation during data collection. Ultimately, this pipeline demonstrates how modern data science can bridge the gap between raw feedback and confident, data-driven decision-making.