DATA3001 FINAL REPORT

SELF-AWARENESS AS A PREDICTOR OF JOB CANDIDATE FIT

BY ecruitalytics

Charmaine Leow (z5255005)
Forest Yang (z5255664)
Lehan Zhang (z5209986)
Su Lin Cheah (z5290577)
Wenxuan Zou (z5212728)

Submitted in partial fulfilment of the requirements of the capstone course DATA3001

Self-Awareness as a Predictor of Job Candidate Fit

Recruitalytics

Our mission is to develop an innovative data-driven alternative to manual CV-screening. We utilize candidate self-awareness to propel a streamlined and unbiased recruitment process that provides an unparalleled experience for both recruiters and applicants.

Executive Summary

Recent developments in data science and machine learning have led to the consideration of data-driven solutions and advancements to many traditional sectors such as job recruitment. Given the time and resource intensive processes currently involved in recruitment, our client Alooba has sought our services to optimise and improve on their recruitment products and revolutionize how companies approach recruitment. This project has identified that candidate self-awareness is of statistical significance to predicting strong technical test performance and perceived aptitude for the role, so it should be taken into account when recruiters assess candidate fit. We establish a suitable quantitative definition of self-awareness and find that self-awareness differs between demographic groups. In particular, we find gender, age, years of experience, location and curriculum vitae (CV) relevance to have significant effects on both predicting test performance and self-awareness. Our linear model demonstrates a clear positive relationship between self-awareness (measured by an average of self-rating on proficiency of tested skills) and candidate test performance when considered alongside candidate characteristics and CV features. Our classification model provides parsimonious predictions of whether a candidate is self-aware or not. We provide suggestions on how Alooba can integrate our findings into their product development and considerations for how to best extract value out of the data they collect to continue innovating and developing their products into the future.

All relevant code and datasets can be found in the folder below. Refer to relevant sections of the report for corresponding code and datasets used.
https://drive.google.com/drive/folders/1YHUEO052JwILgFWPDfCeaq_NuHfgv0AF?usp=sharing

1. Background

Alooba pitches their products as a replacement for the non-standardized manual curriculum vitae (CV) screening procedure. Accordingly, important recruitment process dimensions such as efficient process time (reduced time-to-hire), process quality (identifying the best candidates) and stakeholder satisfaction must be upheld (Laumer et.al., 2015). In conjunction with collecting candidate demographic information and their performance in role relevant skill assessments, Alooba also asks job candidates to provide a self-rating of their confidence in tested skills prior to attempting the test. We intend to leverage this data and develop a systematized and unbiased recruitment methodology for Alooba to integrate into their products and market as a replacement for manual CV screening.

Manual CV screening is used to determine the suitability of a candidate through identification of dispositional attributes which signal a strong fit for the role. Attributes include capacity to meet demands of the role and alignment to organisational values (Tsai, 2010). However, recruiters do not follow standardized patterns and criteria for assessment so their individual experiences and background can bias selection. This can manifest in positive 'similar-to-me' effects or adversely, the horn effect (Nisbett et. al., 1977). This recruiter bias is eradicated by our model which considers established recruitment variables like relevant experience and education alongside self-awareness to identify candidate fit.

Candidates who possess strong self-awareness, or ‘the ability to see oneself clearly’ (Harvard Business Review, 2021) and also demonstrate proficiency in relevant skills are likely to be a great fit for the role. Yet, existing literature suggests the act of self-reflection (through self-rating) can impact task performance contingent on expectation of success and self-belief. This can manifest as discrepancies in self-awareness between groups where the candidates share similar characteristics e.g. gender or experience (Miksch, 2018). No prior research on the efficacy of using self-awareness to assess candidate fit exists.

Further studies have shown differences in groups of candidates (we take gender and individual experience as an example here) performance produce systematically higher or lower assessment scores. A potential explanation of this disparity is that males and females have different test-completion strategies (Balart and Oosterveen, 2019). These strategies can be defined as any reason to lead candidates to answer questions in an order different from the order being administered such as reviewing all answers before submission. Individual experiences like living abroad can increase self-awareness and lead to candidates making clearer career decisions (Charlton, 2018). Research finds that living in another country away from home encourages self-reflection while grappling with cultural differences in a new location, ultimately promoting greater self-awareness and clarity. Another aspect worth noting is that self-rating by individual candidates are also often biased or inaccurate as they may strive to present themselves favourably according to current cultural norms and expectations (Dodd-McCue and Tartaglia, 2010). This form of self-reported response bias serves as a poor reflection of self-awareness. Work-sample tests that resemble tasks candidates will be doing in the job are good indicators of future job performance since skills assessments similar to those provided by Alooba forces employers to critique the quality of a candidate’s work versus unconsciously judging them based on appearance, gender, age, and even personality which can eliminate selection bias (Knight, 2017).

Recent studies have affirmed the value of pre-employment testing (Criteria Corp., 2021). However, employers should not completely discard their intuition in favour of machine algorithms but also gather job-related information like work experiences and qualifications which cannot be measured by pre-employment tests. Incorporating additional insights such as these would have to be methodical in order to avoid being misled by individual perceptions unrelated to the job scope. Skills assessment tests provide predictive validity when compared to traditional means of obtaining information on candidates but fundamentally should only be used as a portion of the recruitment process of potential candidates. These tests are advantageous in standardizing data to aid employers in making more informed decisions in the hiring process. Thus, we draw on both mathematical error functions and psychological research to define an objective quantitative measure of self-awareness to be used as a factor for consideration in predicting strong candidates.

1.1 Objectives

We aim to establish the relevance of self-awareness in determining candidate fit for a role and propose a parsimonious model for defining self-awareness by testing the following hypothesis:

1. We expect differing self-awareness between demographic groups as identified by relevant explanatory variables extracted from pre-test survey information and candidate CVs.
2. Candidate self-awareness can be quantitatively measured and is a useful predictor in identifying strong candidates, as measured by technical test performance.
3. It is possible to quantitatively define a function for self-awareness and build a valid and representative classification model for predicting whether a candidate is self-aware or not.

2. Scope and Project Approach

2.1 Deliverables

The target outputs of our project are a set of strong predictive and classification models supported by extensive descriptive analysis that address our objectives outlined above. We have delivered a predictive model which uses self-rating (as a proxy for self-awareness) to forecast candidate test performance and establish the relevance of considering self-awareness in candidate recruitment, and a classification model which categorizes candidates as either self-aware or not. We explore alternative methods of measuring self-awareness and consider what key characteristics are integral in determining self-awareness. These findings can be integrated into Alooba's products, enabling them to offer their clients an innovative recruitment methodology. Key deliverables in this project include a presentation which summarized our research aims and objectives and this report. The presentation emphasized how our findings align with Alooba's purpose and highlighted our process and key value proposition for Alooba. This 30 page report documents the process and conclusions drawn from exploratory data analysis (EDA), model selection, statistical analysis and visualizations will be provided.

2.2 Plan

We followed our initial timeline proposed in the proposal effectively and delivered our presentation which documented our approaches to the project and its business value to Alooba and are on track to finalize our final report within the time frame proposed. However throughout the process of completing the project, we identified a number of issues that arose and dealt with them accordingly as outlined below.

We approached the project by developing a firm understanding of the client's goals and expectations of the project to ensure we fully understood the problem specifications. We established a research question that we explored by analysing existing research and potential definitions of self-awareness. This was our initial approach to the project and the stage at which we assigned tasks among each team member. We also sought early clarification from our project advisor and the client during the first industry partner visit to avoid any misinterpretation about the problem.

Following the third and fourth week of the project, we used preliminary analysis of the provided datasets (Combined Results and Senior Commercial Analyst) in conjunction with research findings to construct hypotheses and investigate predictive modelling methods appropriate to each dataset. We found data issues such as the failure to declassify sensitive candidate information and misalignment of test scores, inconsistency with *candidate_id* across different datasets and missing information. We promptly raised these issues with our supervisor and the client so that they could be fixed. We drafted and finalized our project proposal for submission during week 4.

Week 4 to 7 of the project proved to be the most challenging since we were constantly faced with issues of data discrepancies and had to resolve them during consultations with the industry partner and project advisor amidst the cleaning and processing of the datasets. Due to the time constraints, we devoted the majority of our time to finalizing our regression model and the metrics used to determine self-awareness, while including a visual descriptive analysis for each of the 3 proposed hypotheses, also adhering to our initial plan to the greatest extent possible. This critical milestone made us reassess our approach to the project which consequently led to further discussion of our methods on handling missing values in our datasets and regression methods appropriate to our model. Additionally, we redistributed the tasks among ourselves according to our individual strengths to further optimize our time. Here, Forest, Charmaine and Lehan focused on model-building (selecting the most representable predictive model to include), fine-tuning the parameters for models in each of the hypothesis and trying to build a good model, as well as elaborating on each empirical strategy adopted, while Su Lin and Wenxuan assisted in reorganizing and summarizing the findings and predictive model, expanding our understanding of the research question through further research and documenting the conclusions. We thus successfully completed the draft report by the week 8 deadline.

In weeks 9 and 10 we focused on improving our predictive models and presentation content based on the feedback obtained on our draft report. We incorporated these in our presentation, which was subsequently showcased among other groups during week 10. Refer to appendix 2.1 to see point-by-point how we responded to the corresponding feedback.

Through this project we have honed our skills in research and data analysis. With our increased understanding of the client and its business, and are now capable of producing deliverables resembling a data science consultancy but also acknowledge that unknown challenges will remain for all upcoming projects in the future.

3. Data

To extract maximum value from the data provided by the client, we have used both datasets provided for the Candidate Self-Awareness project and data from the CV and test performance project. We have complimented this data with external datasets about countries and university ranking to draw conclusions about the value and relevance of variables.

Our preliminary dataset consists of 36 variables extracted from '*Combined Results.xlsx', 'Senior Commercial Analyst (Amsterdam).xlsx', 'DataCountry.dta', 'QS World University Rankings.csv'* and '*FinalCVs_18102021.zip'.* A brief data description of these data sets and variables of interest used in the project is provided below. All data sets are included in our code folder on google drive, link is after the executive summary.

Table 1: A description of the data sets that our variables originated from

Data Set	Description and Notes
'Combined Results.xlsx' *	This dataset contains the results of several assessments on Alooba for a variety of different roles with a breakdown of scores by skill and self-rating. Candidate pre-assessment questionnaire answers which include demographic information such as date of birth, location and years of experience. N=965 observations spanning 7 unique tests (*test_id)* which cover similar skills. Assumptions made include: the skills tested are consistent across the same *test_id,* evidence of attempting the test provided by the presence of both '*Parts Completed'* =3 and a self-rating score (assumed to be a requirement in the pre-test survey). To extract more background information about the candidates, we chose to only include a subset of this data who had a CV- those with *test_id*=1165, candidates we identified to be a subset of the *'Senior Commercial Analyst.xlsx'* dataset. This subset included N=282 samples, some of which were missing variables. Exploration into missing values present in this dataset can be found in Appendix 3.1.
*'Senior Commercial Analyst (Amsterdam) .xlsx'**	This dataset consisted of N=1128 observations, of which N=504 had corresponding CVs provided. Using the variable Candidate_ID, we identified that N=275 of these candidates with CVs had corresponding pre-test questionnaire answers from 'Combined Results.xlsx' . These N=275 unique observations as identified by Candidate_ID are the final population we have chosen to use for modelling.
'*FinalCVs_ 18102021.zip '**	N=504 .doc CV files were provided by the client. We isolated the N=275 files corresponding to the candidates of interest and converted them to .pdf via Adobe Acrobat and extracted the data to .txt via R. A document corpus was created for natural language processing. Preliminary analysis into common words, themes (through topic modelling) and associations present across all the CV data was conducted, however the relevance of this information to address our hypothesis was deduced to be limited. Instead, we extracted variables mimicking Applicant Tracking Software (ATS) quantification of candidate fit and CV relevance through bag of words (BOW) analysis with relevance to the skills tested by the Alooba technical test. Refer to 2.2 Data Description and Appendix 3.3 and 3.4 for how these variables are defined. We also extracted the university a candidate attended, highest level of education obtained, highest company position and technical and soft skills.

*'DataCountry.dta'***	This dataset provided by Miguel Lorca was used to group *Location* provided by *'Combined Results.xlsx'* into Continents. We had originally matched to '*Countries-Continents.csv*'^ and classified countries using dummy variables, however due to the distribution of countries which applicants were from, we chose to isolate *United Arab Emirates* and *India* due to high observations corresponding and this code was already provided by Miguel Lorca in building '*DataCountry.dta'*. This breakdown of countries is provided in Appendix 3.2.
'QS World University Rankings.csv'	This dataset provided an incomplete sample of QS World University rankings for 2020,2019 and 2018. 2020 had the most observations, so it was used alongside the QS Top Universities online database^^ to match candidate universities (extracted from CVs) to their corresponding QS 2020 ranking

**Provided by the client, Alooba*

*** Provided by Miguel Lorca*

^ *'Countries-Continents.csv' from* https://github.com/dbouquin/IS_608/blob/master/NanosatDB_munging/Countries-Continents.csv

^^ *QS TopUniversities Online Database from* https://www.topuniversities.com/search

3.1 Data Description

Our preliminary dataset was used across all our models to ensure consistency of input, however as our models were used for different purposes to address separate hypotheses, the variables included in each model were not the same. Hypothesis testing was undertaken to address hypothesis 1 and determine if it was reasonable to exclude certain variables that did not contribute to identifying differing self-awareness between groups. Refer to sections 4.1 and 5.1 for more detail on variable selection.

Here we will briefly introduce all the variables used across our models and detail how we constructed the variables that were not directly found in data sets.

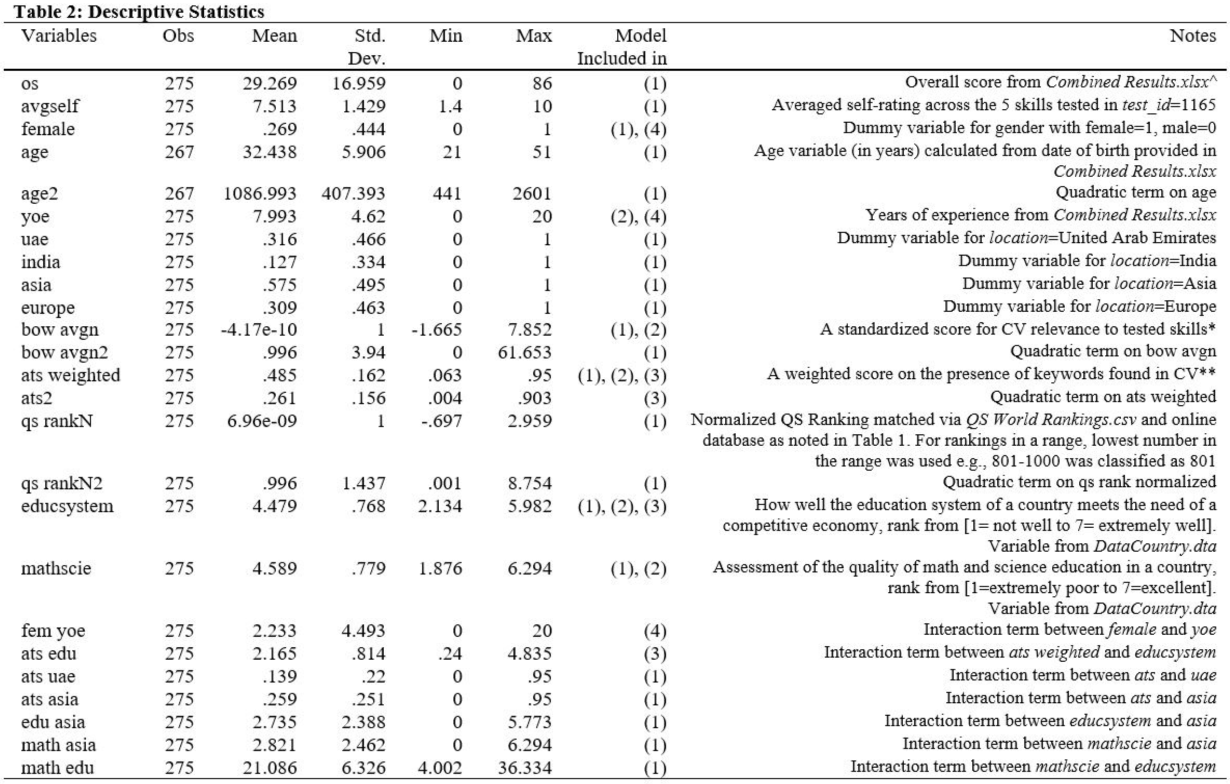


Table 2: Descriptive Statistics

Variables	Obs	Mean	Std. Dev.	Min	Max	Model Included in	Notes
os	275	29.269	16.959	0	86	(1)	Overall score from *Combined Results.xlsx*^
avgself	275	7.513	1.429	1.4	10	(1)	Averaged self-rating across the 5 skills tested in *test_id*=1165
female	275	.269	.444	0	1	(1), (4)	Dummy variable for gender with female=1, male=0
age	267	32.438	5.906	21	51	(1)	Age variable (in years) calculated from date of birth provided in *Combined Results.xlsx*
age2	267	1086.993	407.393	441	2601	(1)	Quadratic term on age
yoe	275	7.993	4.62	0	20	(2), (4)	Years of experience from *Combined Results.xlsx*
uae	275	.316	.466	0	1	(1)	Dummy variable for *location*=United Arab Emirates
india	275	.127	.334	0	1	(1)	Dummy variable for *location*=India
asia	275	.575	.495	0	1	(1)	Dummy variable for *location*=Asia
europe	275	.309	.463	0	1	(1)	Dummy variable for *location*=Europe
bow avgn	275	-4.17e-10	1	-1.665	7.852	(1), (2)	A standardized score for CV relevance to tested skills*
bow avgn2	275	.996	3.94	0	61.653	(1)	Quadratic term on bow avgn
ats weighted	275	.485	.162	.063	.95	(1), (2), (3)	A weighted score on the presence of keywords found in CV**
ats2	275	.261	.156	.004	.903	(3)	Quadratic term on ats weighted
qs rankN	275	6.96e-09	1	-.697	2.959	(1)	Normalized QS Ranking matched via *QS World Rankings.csv* and online database as noted in Table 1. For rankings in a range, lowest number in the range was used e.g., 801-1000 was classified as 801
qs rankN2	275	.996	1.437	.001	8.754	(1)	Quadratic term on qs rank normalized
educsystem	275	4.479	.768	2.134	5.982	(1), (2), (3)	How well the education system of a country meets the need of a competitive economy, rank from [1= not well to 7= extremely well]. Variable from *DataCountry.dta*
mathscie	275	4.589	.779	1.876	6.294	(1), (2)	Assessment of the quality of math and science education in a country, rank from [1=extremely poor to 7=excellent]. Variable from *DataCountry.dta*
fem yoe	275	2.233	4.493	0	20	(4)	Interaction term between *female* and *yoe*
ats edu	275	2.165	.814	.24	4.835	(3)	Interaction term between *ats weighted* and *educsystem*
ats uae	275	.139	.22	0	.95	(1)	Interaction term between *ats* and *uae*
ats asia	275	.259	.251	0	.95	(1)	Interaction term between *ats* and *asia*
edu asia	275	2.735	2.388	0	5.773	(1)	Interaction term between *educsystem* and *asia*
math asia	275	2.821	2.462	0	6.294	(1)	Interaction term between *mathscie* and *asia*
math edu	275	21.086	6.326	4.002	36.334	(1)	Interaction term between *mathscie* and *educsystem*

Model (1) refers to our OLS Linear Regression (see equation (2)). Model (2) refers to our classification self-awareness classification model from method 1 (see equation (3)). Model (3) and (4) refer to the specifications for Method 2 classification (see table 7).

^ Overall score *os* is not an average of the individual skill test results but a weighted score determined by the client. We were unable to obtain information about how this score was weighted but we understand that it is representative of the performance of candidates across all tests and is comparable to the use of individual skills test scores for defining self-awareness in our classification models (which require the individual test scores and associated self-ratings).

*Applicant Tracking Software (ATS) is a popular recruitment product used by companies currently as a replacement to manual CV screening (Shields, 2018). ATS works by assigning points given to specific keywords (determined by the company advertising the position) as a percentage ranking of how many of the keywords appear in a candidate's resume. These keywords are commonly sourced from the job description of the advertised role. We have chosen to follow this method, selecting both "must have" skills and "nice have" skills noted in the job description provided by the client. See: https://apply.workable.com/alooba/j/D9ADAE5522/ Refer to Appendix 3.3 for keywords and method of weighting.

**Bag of words (BOW) is a common natural language processing (NLP) technique in text modelling to extract features from text data and determine relevance (Mujtaba, 2020). We have used a simplified approach of BOW to determine CV relevance to the skills tested by Alooba. Some of these skills are similar to those identified by the company advertising the position such as sql, however many of these skills and their attributes are not known to the candidate. We chose to include this variable to measure innate alignment and fit of candidates via the predisposition of possessing relevant and suited skills which can be useful in capturing innate contributors to self-awareness. Refer to Appendix 3.4 for further information on how this variable was constructed.

For models (2),(3) and (4), the data set used included all n=275 observations found in our preliminary dataset. For model (1), a subset of n=267 observations were included, to ensure that all observations have a corresponding age value. We chose not to extrapolate the missing age value for the 8 missing observations to preserve the representativeness of our analysis.

We note that *asia* includes the observations that have a *uae*=1 and *india*=1 value. We chose to separate *uae* and *india* as these countries had a significantly higher number of observations (refer to Appendix 3.2) We chose not to exclude these observations from the dummy variable *asia* to provide a representative variable covering candidates from Asian countries as there were few candidates from other Asian countries. To ensure that multicollinearity was not an issue, variance inflation factor (VIF) was used to identify the strength of correlation between these variables. We obtain a VIF score between 1 and 5 for each relevant explanatory variable so we deduce that this moderate correlation is not severe enough to require attention and our parameterisations are suitable for use in model (1).

4. Empirical Strategy

This section describes the approach developed to address our three hypotheses which are used to establish the relevance of self-awareness in determining candidate fit for a role and predicting self-aware candidates.

4.1 Defining Self-Awareness

Self-awareness is an inherent trait that one possesses so it cannot be physically measured. Common practice supported by managerial psychology literature is to measure self-awareness through 360-degree feedback systems (Fletcher and Bailey, 2003). The extent to which one's self-rating of behavioural dimensions or competencies are congruent with ratings of the individual by others provides a measure of the degree to which individuals understand their own strengths and weaknesses. The data provided by Alooba includes a self-rating of a candidate's confidence in tested skills. One could argue that this does not adequately capture behavioural factors of self-awareness such as assertiveness or resilience but given our data limitations we believe that integrating this self-rating into our definition provides the most valuable measure of candidate self-awareness for Alooba, given the data they currently collect.

For our improved definition and measure of self-awareness, we chose to look at the difference between perceived performance from self-rating and actual test performance as a measure of self-awareness. We chose to quantitatively define self-awareness using the following equation**:**

$$Self\ Awareness_i = \frac{1}{number\_of(j)} \sum_i (Test\ Score_{ij} - Self\ Rating_{ij})\ where\ j \in \{skills\}\ and\ Self\ Rating_{ij} \neq 0 \qquad (1)$$

This formula gives each candidate a self-awareness score by taking the average of differences between test score and self-rating across skills identified by the presence of self-rating. The Self Awareness score ranges from negative to positive, allowing identification of the direction of bias and establishing group effects. Since we have converted skill attributes for each candidate to feature columns, each candidate now has many skills in which they have no score or self-rating. The standardized nature of this score allows direct comparison between candidates who have either taken tests with a differing number of skills or self-rated only a subset of all skills tested. Should Alooba wish to add an additional test or change the skills tested in the skills test, this score will still be valid as it has weights all skills equally.

We use this quantitative definition in addressing both hypothesis 1 and 3 to establish group differences in self-awareness and to predict and classify candidate self-awareness. In future, Alooba should consider integrating additional measures of self-awareness that do not account for self-rating or can be used in conjunction with self-rating such as using psychometric tests (Fletcher and Bailey, 2003).

4.2 Hypothesis 1- Identifying group differences in self-awareness

Our first hypothesis is used to identify demographic group differences in self-awareness. As identified in the literature review, factors correlated with personal self-belief and expectations can lead to discrepancies in self-awareness. We previously identified that self-awareness could be contingent on gender, location and age/years of experience so we conducted hypothesis testing to determine if mean self-awareness differed within these categorical groups. For completeness, we also conducted hypothesis testing for other variables in the provided datasets and scraped from candidate CVs.

4.2.1 Hypothesis Testing on variables extracted from candidate CVs

One of our goals in this project is to address Alooba's objective which is to provide alternative solutions to manual CV screening. We would like to determine whether certain variables provided in the CVs are useful in measuring candidates' self-awareness, and if so, instead of manually screening CVs, Alooba could expand their pre-test questionnaire to include these variables or use automated screening algorithms to extract them from the CV.

Multiple features were extracted from CVs through scraping. These include QS rank, highest degree level, highest company positions, education systems, math & science education, ATS_weighted scores and BOW scores. Description of these variables can be found in section 3.1 and distributions of some of these variables can be found in Appendix 4.1. Different tests were used to conduct our hypothesis testing to determine if there exist differing self-awareness between variable groups. These significant variables will then be considered when fitting our predictive and classification models. The specific test used to perform hypothesis testing will be dependent on whether variances are equal between groups (Bartlett's test) and if they are normally distributed (Shapiro-Wilk's test). ANOVA test and Student's t-test used in this section requires normality assumption and equal-variances to be satisfied. All hypothesis tests will be evaluated using the 5% level of significance.

Table 3: Summary of hypothesis testing methods for CV extracted variables

Variables	**Groups**	**Normality & Equal Variances**	**Hypothesis Test**
QS rank	None, High Ranking Universities, Low Ranking Universities	Yes, Yes	**ANOVA test** *H0: means of Self-Awareness are equal for different QS rank vs* *H1: means of Self-Awareness are not equal for different QS rank*
Highest degree level	Bachelor, Master	Yes, Yes	**Student's t-test** *H0: means of Self-Awareness are equal for both degree vs* *H1: means of Self-Awareness are not equal for both degree*
Highest company position	graduates/non-senior roles, non-manager/senior roles, manager low levels, senior manager/executive roles	Yes, Yes	**ANOVA test** H*0: means of Self-Awareness are equal for all categories of job positions vs* *H1: means of Self-Awareness are not equal for all categories of job positions*
Education System	1 (poor) – 7 (outstanding)	Yes, Yes	**ANOVA test** H*0: means of Self-Awareness are equal for all quality of education systems vs H1: means of Self-Awareness are not equal for all quality of education systems*
Math & Science Education	1 (poor) – 7 (outstanding)	Yes, Yes	**ANOVA test** H*0: means of Self-Awareness are equal for all quality of maths and science educations vs H1: means of Self-Awareness are not equal for all quality of maths and science educations*

4.2.2 Hypothesis Testing on Pre-Test Questionnaire Variables

Hypothesis testing on the self-awareness scores in pre-test survey information was done in a similar fashion to the variables extracted from CVs. The parametric test ANOVA was used where variance and normality conditions for each group were satisfied, and the non-parametric equivalent of the T-test, the Mann-Whitney U test, was used where the normality and equality of variance assumptions were violated. Similar to the aforementioned, these tests will use a significance level of alpha = 0.05.

Multicollinearity between variables was tested using the Variance Inflation Factor (VIF), which measures the correlation and strength of association between the explanatory variables in a regression model (Frost, 2021). The value of the VIF indicates if multicollinearity exists in the model.

Table 3: Summary of hypothesis testing methods for Pre-Test Questionnaire Variables

Variable	**Groups**	**Hypothesis Test**
Gender (Senior Analyst)	Female, Male	**Mann-Whitney U test** H_0: Populations of the two groups equal H_A: Populations of the two groups not equal
Location (Senior Analyst)	Asia, Europe, Africa	**ANOVA** H_0: Populations means of groups equal H_A: Populations means of groups not equal

Years of Experience (Senior Analyst)	0-5, 5-10, 10-15, 15-20, 20+	**ANOVA** H_0: Populations means of groups equal H_A: Populations means of groups not equal
Gender (Combined Results)	Female, Male	**Mann-Whitney U test** H_0: Populations of the two groups equal H_A: Populations of the two groups not equal
Location (Combined Results)	Australia, Canada, Iran etc. n=20 locations	**ANOVA** H_0: Populations means of groups equal H_A: Populations means of groups not equal
Years of Experience (Combined Results)	0-5, 5-10, 10-15, 15-20, 20+	**ANOVA** H_0: Populations means of groups equal H_A: Populations means of groups not equal
Age & Years of Experience	Linear/Continuous	**Variance Inflation Factor (VIF)** Low score: No multicollinearity High score: Multicollinearity exists

4.2 Hypothesis 2- Self-awareness relevance

The client Alooba is interested in determining the value of considering self-awareness in recruitment. As their products currently focus on the pre-interview stages of recruitment, it is sensible to assume that the measure of success for a good candidate is progressing through to interviews upon completing the Alooba technical skills testing.

To establish the relevance of self-awareness as a measure of strong candidates, we use the overall score achieved by a candidate in the Alooba online assessment as a proxy for strong candidates. We make a crucial assumption that Alooba's testing is valid and test performance reflects a candidate's understanding of relevant skills. If the skills tested are reflective of skills used on the job, we can assume that higher test scores will correlate to higher success in job performance, however if the test cannot assess the ability to perform the job, it is not useful in the selection process and could lead to inappropriate candidates progressing to interviews (Gusdorf, 2008).

As overall score is a weighted average across candidate performance in individual skill tests, we cannot use our quantitative definition of self-awareness as an explanatory variable of overall score due to endogeneity. Thus, we use the average of self-rating across tested skills reported by candidates in the pre-test questionnaire as a proxy for self-awareness as it is the only relevant metric collected by Alooba. We note that self-rating is a reliable indication of how much confidence one has about their own abilities, but it can be biased by candidate characteristics such as gender and experience. These variables and relevant interaction terms are included in our model to isolate the effect of self-awareness on test performance ceteris paribus.

We choose a OLS linear regression specification for our model of the relationship between overall score and self-rating, characterized by equation (2) below:

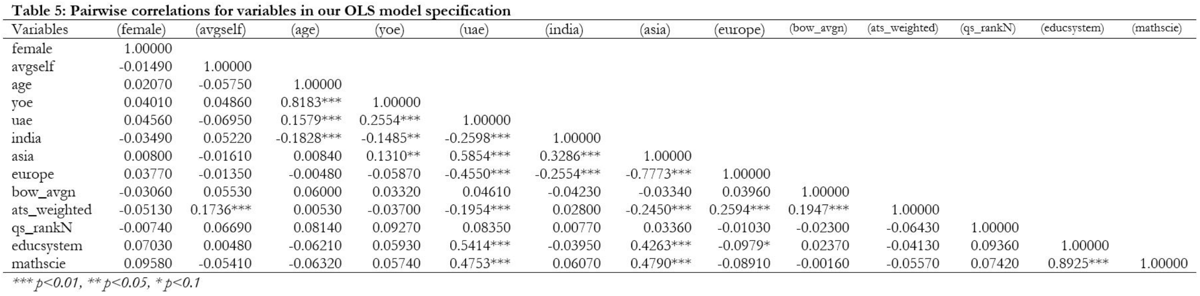
$$\begin{aligned} overall\ score = {} & \beta_0 + \beta_1 female + \beta_2 avgself + \beta_3 age + \beta_4 age^2 + \beta_5 uae + \beta_6 india + \beta_7 asia \\ & + \beta_8 europe + \beta_9 bow\_avgN + \beta_{10} bow\_avgN^2 + \beta_{11} ats\_weighted + \beta_{12} qs\_rankN \\ & + \beta_{13} qs\_rankN^2 + \beta_{14} educsystem + \beta_{15} mathscie + \beta_{16} ats\_uae + \beta_{17} ats\_asia \\ & + \beta_{18} edu\_asia + \beta_{19} math\_asia + \beta_{20} math\_educ \end{aligned} \qquad (2)$$

OLS is an appropriate specification for this model because our dependent variable *overall score* is continuous and normally distributed (refer to Appendix 4.2 for distribution). We justify that our model is correctly set up through the testing methodology specified below.

First, we performed heteroscedasticity tests on the original OLS linear regression. In econometrics, a common test for heteroskedasticity is the White test, which allows the heteroscedasticity process to be a function of one or more independent variables (White, 1980). It is similar to the Breusch-Pagan test, but allows for non-linear and interactive effects of the independent variables on the error variance. We propose that some explanatory variables may have a non-linear relationship with the dependent variable, thus we choose to use the White test. We conducted a White test for heteroscedasticity with null hypothesis H0: Constant variance among residuals and achieved a p-value of 0.0395 thus rejecting the null hypothesis and conclude that heteroscedasticity is present in the data. Hence, we proceed to fit a linear regression model with robust standard errors.

We use the Pearson correlation test statistic to measure the statistical relationship or association between two variables. It is based on the covariance method and is considered to be the best method for measuring associations between variables of interest (Glen, 2021). It provides information about the magnitude of association or correlation, as well as the direction of the relationship.

Table 5: Pairwise correlations for variables in our OLS model specification



Variables	(female)	(avgself)	(age)	(yoe)	(uae)	(india)	(asia)	(europe)	(bow_avgn)	(ats_weighted)	(qs_rankN)	(educsystem)	(mathscie)
female	1.00000												
avgself	-0.01490	1.00000											
age	0.02070	-0.05750	1.00000										
yoe	0.04010	0.04860	0.8183***	1.00000									
uae	0.04560	-0.06950	0.1579***	0.2554***	1.00000								
india	-0.03490	0.05220	-0.1828***	-0.1485**	-0.2598***	1.00000							
asia	0.00800	-0.01610	0.00840	0.1310**	0.5854***	0.3286***	1.00000						
europe	0.03770	-0.01350	-0.00480	-0.05870	-0.4550***	-0.2554***	-0.7773***	1.00000					
bow_avgn	-0.03060	0.05530	0.06000	0.03320	0.04610	-0.04230	-0.03340	0.03960	1.00000				
ats_weighted	-0.05130	0.1736***	0.00530	-0.03700	-0.1954***	0.02800	-0.2450***	0.2594***	0.1947***	1.00000			
qs_rankN	-0.00740	0.06690	0.08140	0.09270	0.08350	0.00770	0.03360	-0.01030	-0.02300	-0.06430	1.00000		
educsystem	0.07030	0.00480	-0.06210	0.05930	0.5414***	-0.03950	0.4263***	-0.0979*	0.02370	-0.04130	0.09360	1.00000	
mathscie	0.09580	-0.05410	-0.06320	0.05740	0.4753***	0.06070	0.4790***	-0.08910	-0.00160	-0.05570	0.07420	0.8925***	1.00000

**** p<0.01, ** p<0.05, * p<0.1*

Correlation between coefficients which are statistically significant at the 1% level are marked by three asterisks in the table above. Hence, we selected these correlated variables as the appropriate interaction terms to include in our model.

We also used the multivariate analysis of variance of ANOVA test to analyse which factors have an independent effect on the dependent variable *overall score* and whether the interaction of multiple control factors have a significant effect on the distribution of the observed variables (Kenton, 2021). This ensures that we include the optimal combination that favours the observed variables. We conducted the ANOVA test with null hypothesis: Variables have significant effect. The F-test value for the entire model is 4.28 with a p-value of 0.00001, indicating that it passes the test. The p-values for *female, age, age_sqr, bow_avgn2, qs_rankN, qs_rankN2, educsystem, ats_uae, edu_asia, math_asia, math_edu*, passed the test (p-value< 0.1) so we accept the null hypothesis and conclude that this set of variables have a jointly significant effect on overall score.

We do not find any significant interaction between our explanatory variable of interest *avgself* and other exogenous predictors. This is not of concern as the literature suggests that the interaction between self-rating and other explanatory variables in predicting test performance is inconsistent (Hansford and Hattie, 1982). Wylie (1979) finds that the correlation of achievement indices and overall self-regard indices tend to be small in absolute terms. As we only intend to use self-rating as a proxy for self-awareness here, we acknowledge this limitation in identifying potential interactions between self-awareness and test performance. However, we have firmly established that our aim is not to use self-awareness to purely predict performance in testing but to use it alongside test performance to establish strong candidates so we deem this to be appropriate for our usage.

To determine which independent variables should be included in the established OLS Linear Regression, AIC and BIC were used in model selection criteria. Both these measures explain how well the model will predict on new data. The reason for choosing AIC rather than BIC is that AIC penalizes complex models less, focusing more on model performance (Brownlee, 2019). We chose the model with the smallest AIC score as the most appropriate specification as a lower score means that the model has improved prediction.

To assess the performance of our final OLS model, we use the r-squared score. This is a measure of how well the model explains the observed data by calculating the percentage of variance in the dependent variable that can be explained by all independent variables. The higher the R^2 value, the better the fit of a model so our selected specification should maximize the R^2 score. We expect an R^2 score of between 0.25 to 0.35 for our final model, as it provides confidence that our model explains between 25% to 35% of the total variance of the dependent variable and is in line with literature. Adding more variables decreases predictive accuracy but explains the observed data better, so we use adjusted R^2 to compare against different OLS model specifications with differing independent variables to further support our AIC criteria.

4.3 Hypothesis 3- Predicting self-aware candidates

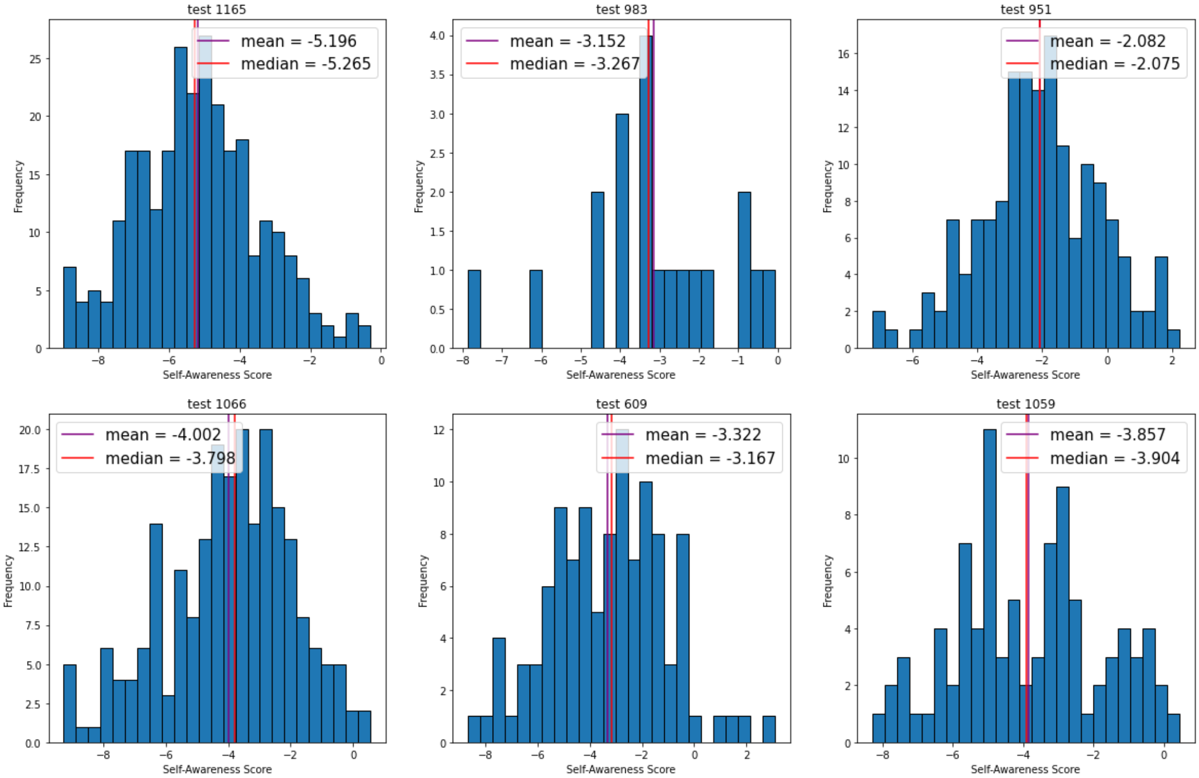
We compare 2 different methods of categorising candidates into groups on whether they are self-aware based on their self-awareness score and producing models to learn and predict if a candidate is self-aware or not. Detailed explanations on how these 2 distinct approaches are set up are provided below.

4.3.1 Method 1

In our first classification method of classifying candidates' self-awareness, we categorise candidates into groups of whether they are self-aware based on the distance of their self-awareness score from 0, taking into account the job position candidates applied for and the difficulty of the test they took. A self-awareness score that is close to 0 indicates that the candidate is "self-aware".

An assumption we made that was verified by Alooba is that the level of difficulty for different tests varies since some tests are designed for a more technical role, and some are more targeted for senior roles. We also validated our assumption by performing descriptive statistics and hypothesis testing and found that the means of self-awareness score are statistically different for different tests. Hence, categorisation of candidates' self-awareness should be different depending on the test they took. A detailed analysis of our hypothesis testing can be found in Appendix 4.3.

Figure 1: Distributions of self-awareness scores for different test_ID



Based on our hypothesis testing results, we produced the following categorisation of candidates' self-awareness for different tests (see Appendix 4.3 for justification of these range):

Table 6: Self-awareness classification for different test_id

	Self-Aware	**Not Self-Aware**
Test_id 951	Self-Awareness score between -1 and 1	Self-Awareness score < -1 or Self-Awareness score > 1
Test_id 609	Self-Awareness score between -2 and 2	Self-Awareness score < -2 or Self-Awareness score > 2
Test_id 983	Self-Awareness score between -2 and 2	Self-Awareness score < -2 or Self-Awareness score > 2
Test_id 1059	Self-Awareness score between -2 and 2	Self-Awareness score < -2 or Self-Awareness score > 2
Test_id 1066	Self-Awareness score between -2 and 2	Self-Awareness score < -2 or Self-Awareness score > 2
Test_id 1165	Self-Awareness score between -3 and 3	Self-Awareness score < -3 or Self-Awareness score > 3

We found that the test for the Senior Commercial Analyst role (test_id 1165) seems to be the most challenging compared to the other 5 tests. Hence, for the Senior Commercial Analyst role, we categorised candidates with a self-awareness score between -3 and 3 as "self-aware".

Based on the variables found that were useful from feature selections in hypothesis 1, and a trial-and-error approach to select the variables that produce the highest ROC_AUC score in our baseline model, we found that the following subset of variables are significant in classifying candidates' self-awareness:

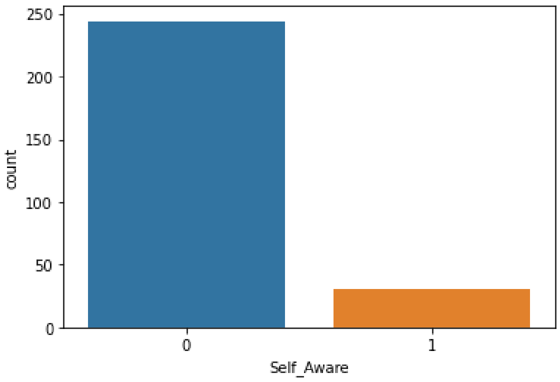
$$Self_{Aware} = \beta_0 + \beta_1 ATS\_weighted + \beta_2 educsystem + \beta_3 mathscie + \beta_4 BOW\_avgN + YoE \qquad (3)$$

We will be focusing in particular using Logistic Regression and Decision Trees algorithms to train and classify our model. Both algorithms are widely used and popular choices for classification problems, with different pros and cons in each. Decision Trees provide better predictability, visualisations and interpretability of results, whereas Logistic Regression captures underlying relationships better with appropriate feature engineering and it is computationally less expensive (Bock, 2021). The success of both algorithms will be evaluated based on how well they fit the data and the ROC_AUC score produced.

ROC_AUC score is a metric used to evaluate the success of all our classification models. The choice of this metric is due to an imbalanced class problem in method 1, nonetheless it is also a great metric for problems with balanced classes. ROC is a probability curve and it is plotted with True Positive Rate against the False Positive Rate at various threshold values. AUC is a measure of separability; it measures how well a model can distinguish between classes. An AUC score > 0.8 is ideal, however a score between 0.7 and 0.8 is considered as an acceptable classifier (Mandrekar, 2015).

A downside of this definition on classifying candidates' self-awareness is that it produces a very high imbalanced class which could be a problem because we run into the risk of algorithms displaying poorer predictive performances. Figure 2 below shows that we have a highly imbalanced class with only 31 out of 275 candidates classified as "self-aware". The minority class, those "not self-aware", has less data for the algorithms to learn from, so the chances of classifiers classifying candidates' self-awareness inaccurately is very high. This would lead to classifying candidates who are “not self-aware” as “self-aware” on unseen dataset.

Figure 2: Self-awareness within test ID=1165 candidates

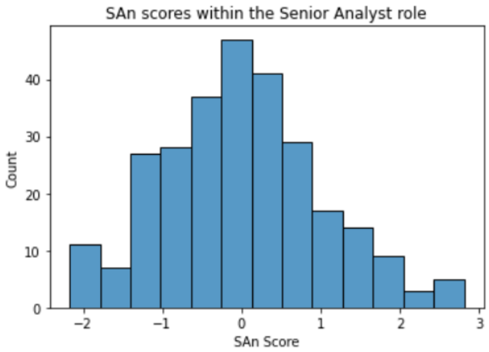


There were a few variables found useful in feature selections that weren’t included in this model because we found that those variables weren’t significant in classifying candidates’ self-awareness through model experimentation. We justify that it is appropriate for this project as we are providing preliminary implementation of self-awareness prediction to Alooba- something innovative and within a new area of research. However, in future research variables found to be significant in feature selections such as gender and location should be incorporated into the model. This would be best practice following econometrics analysis, to explain the effect these variables have on classifying candidates’ self-awareness even if they are not significant.

4.3.2 Method 2

Due to the imbalanced classes problem encountered in Method 1, we now propose two alternative methods of classification, labelled 2a and 2b, both of which will have balanced classes. We first normalize all self-awareness scores within our dataset. We call this new measure “SAn”, short for “Self-Awareness Normalized”. Roughly half of all candidates now have their SAn score more than zero, with the rest less than zero, as shown below. This section will focus on setting up our two main methods of splitting candidates into binary classes using this score.

Figure 3: SAn scores within the senior analyst role



Method 2a

For candidates whose SAn score is above 0, we defined them to be in class “1”, or “self-aware”. Candidates whose SAn score is below 0 are deemed “not self-aware”, and we set their class to “0”. We set classes like this since candidates who are self-aware should have a higher SAn score, which would mean that their self-rated score is close to their actual test score. While this is an assumption, we have shown in the chart above that most candidates will self-rate higher than their actual test score.

Furthermore, class definitions can be swapped in reverse cases, where candidates self-rate lower than what they score in tests.

Method 2b

Similar to method 2a, we utilize a candidate's SAn score. However, now we define someone to be in class "1" or "self-aware" if their SAn score falls within x standard deviations from zero. In this dataset, we choose x as 0.7, though this can be altered for different datasets to create balanced classes. If their SAn score is outside this range, they are classified as class "0". This can be interpreted as a candidate being self-aware if their self-awareness falls around the mean self-awareness for the group in which they are in. To be able to use this interpretation, we assume that the distribution of SAn scores is roughly normal.

A theoretical shortcoming of these definitions is that they both require a substantial amount of pre-existing data from which to draw the classes. This is acceptable for jobs with candidate data available already, or who have had a large number of candidates apply. Furthermore, in this situation, we will only be able to predict whether a future candidate will fall into the upper or lower portion of existing self-awareness scores for all candidates applying for that job.

Feature selection for the variables in the table below was done through testing possible combinations of significant predictors that have been determined through hypothesis testing. Due to the nature of the exponentially increasing number of polynomial and interaction terms, producing a cross-validated test score for every permeable combination was not feasible, computation wise. Thus, features were initially chosen through p-value significance in logistic regression, and then subsequently fitted on the following models. The effectiveness of including interaction terms was then done after the selection of important linear terms.

Table 7: Models tested and learning algorithms used in Method 2

	Learning Algorithms	**Variables Included**
Method 2a	Logistic Regression	$\beta_0 + \beta_1 EducSystem + \beta_2 Age + \beta_3 ATS_{Weighted} + \beta_4 EducSystem \times ATS_{Weighted} + \beta_5 ATS_{Weighted}^2$
	Perceptron	$All\ Features\ Included$
	Adaboost	$EducSystem + Age + ATS_{Weighted} + EducSystem \times ATS_{Weighted} + ATS_{Weighted}^2$
	Decision Trees	$EducSystem + Age + ATS_{Weighted} + EducSystem \times ATS_{Weighted} + ATS_{Weighted}^2$
Method 2b	Logistic Regression	$\beta_0 + \beta_1 Female + \beta_2 YearOfExperience + \beta_3 Female \times YearsOfExperience$
	Adaboost	$All\ Features\ Included$
	Decision Trees	$EducSystem + Age + ATS_{Weighted} + EducSystem \times ATS_{Weighted} + ATS_{Weighted}^2$

The Perceptron and Adaboost algorithms were selectively chosen in addition to the Logistic Regression and Decision Trees methods used in Method 1 to allow more flexibility in hyperparameter tuning, in case of models with low accuracy. Since Logistic Regression is a specific case of the Perceptron, the exploration of additional activation functions could yield better model results. This methodology also applies for Adaboost. Since Adaboost models consist of Decision Tree stumps, tuning of stump size, the number of stumps and misclassification penalty could produce a better performing model.

Performance of the two methods' models will be evaluated using ROC_AUC scores in a similar manner to Method 1. Due to the balanced classes we have created, we also have the option of using the simpler 'accuracy' metric to compare model performance within these two methods. Cross validation will also be utilised, though this is dependent on our models' computation time.

With more time and further research, the effectiveness of including more interaction terms can be determined. Since there is an exponential increase in computation time with the increase in features (including their interaction terms), we were limited in only selecting and testing interactions between our most significant variables. Furthermore, similar to Method 1, the effect and correctness of the removal of insignificant control variables in some models, such as the removal of gender in Method 2a's Logistic Regression (due to it not being statistically significant), could be investigated in the future.

5. Results

This section describes the results obtained from exploring our 3 hypotheses. We present the EDA results for hypothesis testing and variable selection in section 5.1 (hypothesis 1). In section 5.2 we cover our final OLS linear regression model results and additional model specifications explored to address hypothesis 2. Section 5.3 covers our final classification model specification and analyses the implications of additional parameterisation performance. We present additional findings in section 5.4 that are not directly related to addressing our objectives but we believe bring business value to our client Alooba. All relevant code for reproducing results are in the appendix, on our google drive and uploaded in Moodle.

5.1 Hypothesis 1- Feature Engineering and Descriptive Results

The results presented in this section provide strong support for our hypothesis that self-awareness differs between demographic groups. This section has been separated into examining variables extracted from candidate CV's through NLP methodology in section 5.1.1 and variables from the pre-test questionnaire in section 5.1.2.

More in depth implications on the significance of these variables are provided in sections 5.2-5.4 as they play different roles in providing explanatory power to our different models.

5.1.1 Hypothesis testing on variables from candidate CVs

In this section we will briefly mention the variables we found were significant from hypothesis testing. Results of other variables found to be not significant from hypothesis testing can be found in Appendix 5.1.

Table 8: Results Summary found from hypothesis testing on the CV extracted variables

Variables	**Groups**	**Hypothesis Test**	**P-value**	**Accepted Hypothesis**
QS rank	None, High Ranking Universities, Low Ranking Universities	**ANOVA test**	0.915	*H0: means of Self-Awareness are equal for different QS rank*
Highest degree level	Bachelor, Master	**Student's t-test**	0.024	*H1: means of Self-Awareness are not equal for both degree*
Highest company position	graduates/non-senior roles, non-manager/senior roles, manager low levels, senior	**ANOVA test**	0.418	H*0: means of Self-Awareness are equal for all categories of job positions*

	manager/executive roles			
Education System	1 (poor) – 7 (outstanding)	**ANOVA test**	0.046	*H1: means of Self-Awareness are not equal for all quality of education systems*
Math & Science Education	1 (poor) – 7 (outstanding)	**ANOVA test**	0.218	H0*: means of Self-Awareness are equal for all quality of maths and science educations*

Self-awareness and highest degree level
Under the 5% level of significance, we reject the null hypothesis (p-value = 0.024 < 0.05) and conclude that the means of self-awareness scores are statistically different between bachelor and master degrees. This makes intuitive sense since candidates who obtained a master degree received more education and tend to be more knowledgeable about their skills. We also scraped other categories of degrees such as PhD, however there were only < 10 observations in these, so we excluded them from the testing. Distribution of each degree can be found in Appendix 5.1.

Self-awareness and education systems
Under the 5% level of significance, we reject the null hypothesis (p-value = 0.046 < 0.05) and conclude that the means of self-awareness scores are statistically different between the different quality of education systems received by candidates in their countries. Distribution of each level of education systems can be found in Appendix 5.1. Further analysis into the significance of differing education systems on our OLS model and classification models are explored in sections 5.2 and 5.3 respectively.

5.1.2 Hypothesis testing on pre-test questionnaire results

Below is a chart of results obtained through hypothesis testing on our pre-test survey variables.

Table 9. Results summary found from hypothesis testing on the Pre-test Survey variables

Variable	**Groups**	**Hypothesis Test**	**P-value**	**Accepted Hypothesis**
Gender (Senior Analyst)	Female, Male	**Mann-Whitney U test**	0.002	H_A: Populations of the two groups not equal
Location (Senior Analyst)	Asia, Europe, Africa	**ANOVA**	0.004	H_A: Populations means of groups not equal
Years of Experience (Senior Analyst)	0-5, 5-10, 10-15, 15-20, 20+	**ANOVA**	0.10	H_0: Populations means of groups equal
Gender (Combined Results)	Female, Male	**Mann-Whitney U test**	0.063	H_0: Populations of the two groups equal
Location (Combined Results)	Australia, Canada… Iran	**ANOVA**	5.937×10^{-27}	H_A: Populations means of groups not equal
Years of Experience (Combined Results)	0-5, 5-10, 10-15, 15-20, 20+	**ANOVA**	0.0003	H_A: Populations means of groups not equal
Age, Years of Experience	Linear/Continuous	**Variance Inflation Factor (VIF)**	VIF = 192 (Age)	High score (both): Multicollinearity exists

			VIF = 31 (YoE)	
Highest Role in company	Categorical	**Variance Inflation Factor (VIF)**	N/A	High score: Multicollinearity exists

Self-awareness and gender

We can conclude that there is no difference in mean self-awareness between gender in the Combined Results dataset. On the contrary, when we limit our test to the Senior Analyst position, we discover the opposite is true, that there is a significant difference in mean self-awareness between genders.

The unexpected result in Senior Analyst is likely in part due to the small sample size as well as the possibility that the type of candidates attracted to the Senior Analyst role naturally take these distributions. At first glance of the distributions in the appendix 5.2, we notice that the shape of the distribution of self-awareness for males is slightly higher than that for females, most likely dispelling the common "male overconfidence bias", at least within the Senior Analyst role. Though, in the future, a one-sided T-test can be produced to determine the direction in which the difference is significant. In addition, the low ratio of females to males in both Senior Analyst and Combined Results highlights the gender disparity within the Data Science industry.

Self-awareness and location of applicant

Focusing on the 20 countries with the most candidates, our ANOVA test produces the result that there exists at least one country whose mean self-awareness is not equal to the rest. This can be explained by countries' varying educational systems. These differences are more noticeable in the graph in the appendix 5.2 where we testi the Combined Results dataset. Our results show that the only countries where candidates have a positive self-awareness score are Australia and Canada, both of which are English dominant countries. While it can be considered natural for the English dominant countries to have a higher self-awareness score in a skills test and job posting that is written in English, the distribution of scores of other English dominant countries such as New Zealand and the UK do not wholly support this notion (Major et.al., 2005). However due to the limited sample size we cannot draw strong conclusions about this. When retesting on the distribution within the Senior Analyst role, we group countries into continents instead of countries, due to insufficient sample sizes. However, our ANOVA conclusion remains the same, that there is a difference in the mean self-awareness by continent.

It is important to consider potential cultural biases that may result from considering self-awareness. As identified in the background, cultural norms and expectations can shape how one chooses to present themselves. Research has demonstrated that cultural experience influences cross-situational self-consistency and reliance on positive self-evaluation, both of which could impact how one chooses to rate their confidence on a skill (Lu and Wan, 2018). Alooba should consider implementing additional measures of self-awareness that are culturally aware to ensure that language barriers and societal norms do not lead to English as a Second Language (ESL) job applicants being disadvantaged in the hiring process.

Self-awareness and Years of Experience

Our results reflect the existence of at least one Years of Experience group whose mean self-awareness is different to the others in Combined Results, with reversed results in our Senior Analyst dataset.

The gradual skewness change in the Senior Analyst chart in appendix 5.2 raises the possibility that as candidates spend longer time in the industry, they may be either unaccustomed to taking tests of this nature, or underprepared, as we would normally expect senior candidates to be more accurate in measuring their own performance relative to other candidates and more accurate in deducing their possible test score due to their extended experience within the industry. Though, one counterargument is that since senior candidates are more likely to find a job faster than those with less experience, their lack of practice in timed conditions such as the online tests Alooba provides is revealed (Deursen, Dijk, Peters, 2011). Ensuring that online pre-employment testing is not biased towards a particular set of candidates with certain levels of experience is a crucial consideration for Alooba. This impact of experience on self-awareness and corresponding age on overall score is further explored in section 5.3 and 5.2 respectively.

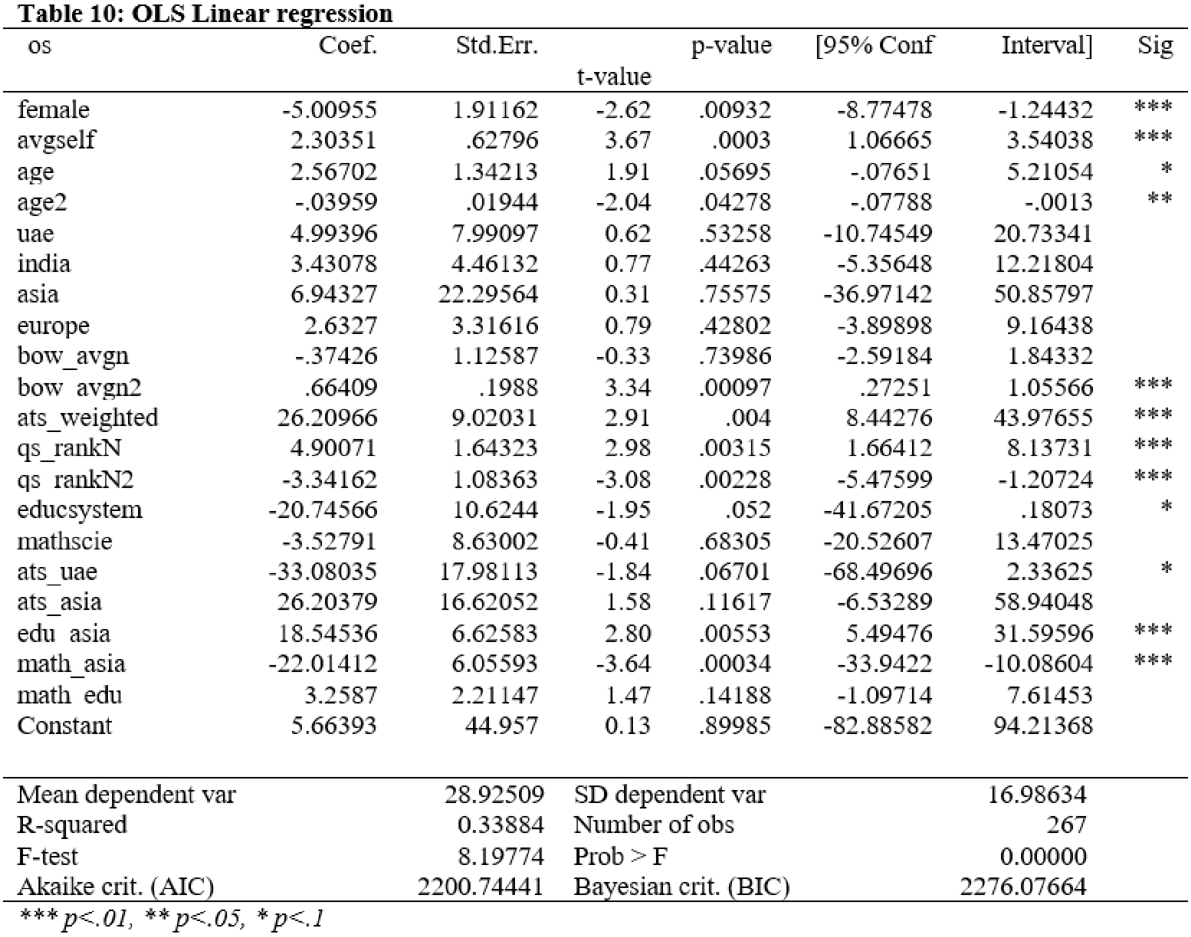
Multicollinearity between Age, Years of Experience and Highest Role in Company

VIF test scores of over 10 show that years of experience, the highest role in a company, and age are all highly collinear. Hence, we chose to include only one of these variables in our model for hypothesis 2 and 3. The variable selected for each model contributed to best model performance based on the respective criteria for success.

5.2 Hypothesis 2- OLS Linear Regression

The final OLS linear regression results are presented in the table below.

Table 10: OLS Linear regression



os	Coef.	Std.Err.	t-value	p-value	[95% Conf	Interval]	Sig
female	-5.00955	1.91162	-2.62	.00932	-8.77478	-1.24432	***
avgself	2.30351	.62796	3.67	.0003	1.06665	3.54038	***
age	2.56702	1.34213	1.91	.05695	-.07651	5.21054	*
age2	-.03959	.01944	-2.04	.04278	-.07788	-.0013	**
uae	4.99396	7.99097	0.62	.53258	-10.74549	20.73341	
india	3.43078	4.46132	0.77	.44263	-5.35648	12.21804	
asia	6.94327	22.29564	0.31	.75575	-36.97142	50.85797	
europe	2.6327	3.31616	0.79	.42802	-3.89898	9.16438	
bow_avgn	-.37426	1.12587	-0.33	.73986	-2.59184	1.84332	
bow avgn2	.66409	.1988	3.34	.00097	.27251	1.05566	***
ats_weighted	26.20966	9.02031	2.91	.004	8.44276	43.97655	***
qs_rankN	4.90071	1.64323	2.98	.00315	1.66412	8.13731	***
qs rankN2	-3.34162	1.08363	-3.08	.00228	-5.47599	-1.20724	***
educsystem	-20.74566	10.6244	-1.95	.052	-41.67205	.18073	*
mathscie	-3.52791	8.63002	-0.41	.68305	-20.52607	13.47025	
ats_uae	-33.08035	17.98113	-1.84	.06701	-68.49696	2.33625	*
ats_asia	26.20379	16.62052	1.58	.11617	-6.53289	58.94048	
edu asia	18.54536	6.62583	2.80	.00553	5.49476	31.59596	***
math_asia	-22.01412	6.05593	-3.64	.00034	-33.9422	-10.08604	***
math edu	3.2587	2.21147	1.47	.14188	-1.09714	7.61453	
Constant	5.66393	44.957	0.13	.89985	-82.88582	94.21368	

Mean dependent var	28.92509	SD dependent var	16.98634
R-squared	0.33884	Number of obs	267
F-test	8.19774	Prob > F	0.00000
Akaike crit. (AIC)	2200.74441	Bayesian crit. (BIC)	2276.07664

**** p<.01, ** p<.05, * p<.1*

We find that the predictor *avgself*, our proxy for self-awareness is statistically significant at a 1% level with a coefficient of 2.304. This can be interpreted as a 2.304 mark increase in overall score given a 1-unit increase in self-awareness, ceteris paribus. Therefore we establish that there is a positive and linearly increasing relationship between *avgself* and *overall score.* This result provides support for hypothesis 2 in establishing that candidate self-awareness is a useful predictor in identifying strong candidates.

Other notable variables of interest include *female, age, bow_avgn, ats_weighted, qs_rankN, educsystem,* and interaction terms for location and education variables which are all statistically significant at the 10% level or above. The implication and interpretation for these significant variables are given below.

We start by noting a coefficient of -5.010 on *female*, our dummy variable for gender. This suggests the presence of gender effects in test performance with females scoring 5.010 marks lower than males ceteris paribus and is in line with literature on gender differences in STEM related assessment under time pressure (Shurchkov, 2012).

Age is also suggested to play a role in determining overall score, with the partial effect of conditional mean with respect to age showing that performance decreases with age. Whether this is due to implicit biases present in accessibility through the format of the test or technological competence, further research is needed to determine whether the testing presents ageism bias and it is an important consideration when designing pre-employment testing and reducing hiring bias (Scepura, 2020).

We find that our variables for CV relevance, *bow_avgn* and *ats_weighted* are both statistically significant at a 1% level. The significance of the quadratic term on *bow_avgn* suggests a positively increasing convex relationship between the presence of relative keywords in the resume and performance in the test. This makes sense as those who list relevant skills on their CV are assumed to possess the relevant experience. We find that *ats_weighted* has a significant impact on overall score, with a 1-mark increase in *ats_weighted* resulting in a 26.210 mark increase in test performance. Candidates who have tailored their resume to the job description as a statement of competency are likely to meet recruiter notions of applicant fit (Bright and Hutton, 2002). Given that these candidates also appear to perform significantly better on the testing, our results suggest that extracting CV relevance is still of great value to the recruitment process and should be replaced with just technical testing alone. We conclude that manual CV screening can be replaced with automated CV screening processes to extract relevant variables that contribute to identifying self-aware candidates and assessing test performance. Further research should be conducted on the value and implications of predicting test performance from candidate CVs as this can also unearth important implications about test taking behaviour, identify inconsistency in performance and ensure that Alooba provides unbiased products.

The high collinearity between location and education variables suggest some interesting relationships between education outcomes and test performance. Care was taken to ensure that the collinearity was accounted for so it did not bias interpretation of our results by conducting a VIF test. Multicollinearity does not affect prediction or goodness of fit however, and we only want to make predictions of the relationship between overall score and self-rating so we add appropriate interaction terms to ensure that the model captures the multicollinear relationships. It is widely accepted that education quality differs between countries around the world, with different education systems focusing on results deemed important to the needs of a competitive economy in their country (Strietholt et. al., 2014). Our results suggest that given students are from Asia, the math and science education results in 25.542 marks lower overall score performance. This is interesting and not in line with literature. The Center for Excellence in Education (CEE) ranks high performing STEM education by the outcomes in International STEM Olympiads, where Asian countries account for 8 of the top 10 highest performing countries (McLean, 2019). However, this discrepancy could be due to the overwhelming representation from UAE and India, neither of whom appear in the top 10. Although we previously identified *ats weighted* to have a large, positive impact on overall score, isolating the impact of *ats weighted* given a candidate is from UAE shows that this positive impact no longer applies. Candidates who are from the UAE score 6.870 marks lower in overall score following a 1-mark increase in *ats weighted.*

Conclusions about a country's education system should not be drawn from our analysis though, as it is not the main focus thus we do not include many variables which may strengthen these conclusions. These results also highlight potential endogeneity within location and education variables, however they were not explored in our analysis as these factors were not the main concern in establishing a relationship between the dependent variable and the explanatory variable of interest, *avgself.* Previously in hypothesis 1, we also identified differences in self-awareness between candidates from different countries. Future research should consider testing for endogeneity in these variables to better understand and mitigate potential biases in location and county of education.

We compared the performance of 6 OLS model specifications to identify which models gave the best performance against our assessment criteria established in section 4.2 and found that model 5 (our final specification) provides the lowest AIC score and a suitable R^2 score.

Table 11. Comparison of Model Performance for OLS Specifications

	Model Specification	**AIC**	**BIC**	**R^2**	**Adjusted R^2**
Model 1	All relevant independent variables but no interaction terms included	2206.737	2271.308	0.3085	0.2613
Model 2	All relevant variables and all non-zero interaction terms	2213.725	2328.517	0.3608	0.2765
Model 3	All relevant variables and all interaction terms statistically significant at a 5% level	2201.767	2284.273	0.3462	0.2872
Model 4	Only independent variables that have an effect on overall score according to ANOVA testing	2247.188	2293.823	0.1646	0.1252
Model 5	**All relevant variables and statistically significant interaction terms, using *age***	**2200.744**	**2276.077**	**0.3388**	**0.2851**
Model 6	All relevant variables and statistically significant interaction terms, using *yoe*	2263.36	2339.312	0.3417	0.2899

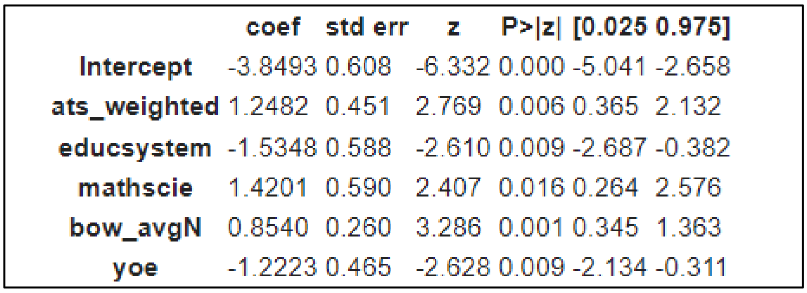
Model 5 provides the smallest AIC score of 2200.744 and also the smallest BIC score amongst all models. The R-squared value of 0.3388 sits between our proposed range of 0.25-0.35 which suggests that the model is suitable at explaining variance on the dependent variable. In comparison to the other models, an adjusted r-squared score of 0.2851 suggests good relative performance in comparison to other models. Thus, we determine that this model is the appropriate specification for establishing the relationship between self-awareness and candidate fit.

5.3 Hypothesis 3- Classification Results

We recall our classification methods Method 1 and Method 2a & 2b from section 4.3. We choose our final classification model to be that of Method 1. The following subsections will show our quantitative results for these methods and discuss the implications and inferential results of the models built.

5.3.1 Method 1

Table 12. Logistic Regression results



	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-3.8493	0.608	-6.332	0.000	-5.041	-2.658
ats_weighted	1.2482	0.451	2.769	0.006	0.365	2.132
educsystem	-1.5348	0.588	-2.610	0.009	-2.687	-0.382
mathscie	1.4201	0.590	2.407	0.016	0.264	2.576
bow_avgN	0.8540	0.260	3.286	0.001	0.345	1.363
yoe	-1.2223	0.465	-2.628	0.009	-2.134	-0.311

Based on the summary output, at 5% level of significance, all predictors used are significant in classifying candidates' self-awareness. We found that ATS_weighted score and the normalised average BOW score have the most effect in categorising candidates' self-awareness. An increase of 1 point in ATS_weighted score multiplies the odds of being "self-aware" by 3.46. Moreover, an increase of 1 point in the normalised average BOW score multiplies the odds of being "self-aware" by 2.35. Both of these results make sense since candidates with more relevant keywords from the job descriptions and skills tested on their CV means that they are knowledgeable on those skills and more likely to perform well in those skills tests.

Literature suggests that the better the quality of the education system an individual receives in their country the better and more successful they will perform in life (Kudroli Foundation, 2019). From our results, going up from 1 level of the quality of education systems to the next divides the odds of being "self-aware" by 0.22. This result is interesting and contradicts the literature, it could be due to candidates who receive better education systems being overconfident with their abilities. However, we found that going up from 1 level of the quality of maths and science education to the next multiples the odds of being "self-aware" by 4.14.

We found that going up from 1 level of Years of Experience to the next divides the odds of being "self-aware" by 0.30. This makes intuitive sense since candidates who have greater experience tend to think that they are very knowledgeable about a specific skill and thus self-rate themselves higher, underestimating the difficulty of the test.

Feature importance score measures how useful a feature is in contributing to the prediction of a model. The score is calculated based on the reduction in Gini, which is the criterion used to select split points in the trees. The sum of feature importances add up to 1 (Brownlee, 2020). Based on the feature importances values produced by the Decision Trees algorithm, the top variables found to be most useful in classifying candidates' self-awareness are ATS_weighted score and the normalised average BOW score, which is consistent with the most effective variables found from Logistic Regression.

Table 13. Feature importances from Decision Trees algorithm

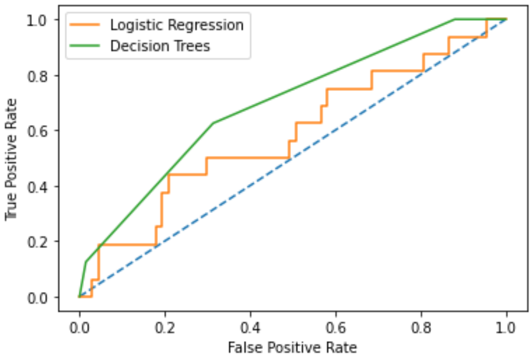
Predictors	Feature importances
ATS_weighted	0.4872
BOW_avgN	0.3112
educsystem	0
mathscie	0
YoE	0.2016

Out of the 2 learning algorithms, Decision Trees performed the best with a testing ROC_AUC score of 0.7, indicating that there is a 70% chance it is able to distinguish between "self-aware" and "not self-aware" candidates. Based on the evaluation threshold we set, it is considered as an acceptable classifier at distinguishing between classes and it is less overfitting compared to Logistic Regression. A higher ROC_AUC score is ideal on the testing set but due to the highly imbalanced classes, both classifiers tend to classify inaccurately on candidates who are "not self-aware", classifying them as "self-aware".

Table 14. Training and testing ROC_AUC score for different algorithms

Learning Algorithms	ROC_AUC Score
Logistic Regression	Training (70%) = 0.8 Testing (30%) = 0.6
Decision Trees	Training (70%) = 0.8 Testing (30%) = 0.7

Figure 4: ROC Curve for learning algorithm



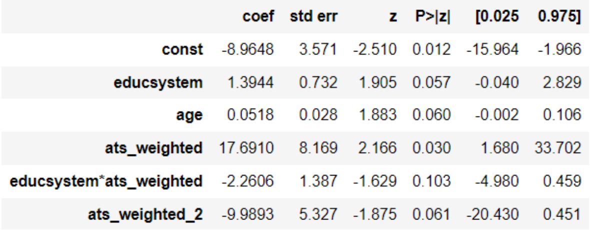
5.3.2 Method 2

Table 15. Summary of results for all models tested

	Learning Algorithms	**Variables Included**	**Accuracy Score**	**ROC_AUC Score**
Method 2a	Logistic Regression	$\beta_0 + \beta_1 EducSystem + \beta_2 Age + \beta_3 ATS_{Weighted} + \beta_4 EducSystem \times ATS_{Weighted} + \beta_5 ATS_{Weighted}^2$	0.6578 (30% Test Set)	0.7150 (30% Test Set)
	Perceptron	$All\ Features$	0.5080 (5-fold CV)	N/A
	Adaboost	$EducSystem + Age + ATS_{Weighted} + EducSystem \times ATS_{Weighted} + ATS_{Weighted}^2$	0.5715 (5-fold CV)	0.5830 (5-fold CV)
	Decision Trees	$EducSystem + Age + ATS_{Weighted} + EducSystem \times ATS_{Weighted} + ATS_{Weighted}^2$	0.5638 (5-fold CV)	0.5254 (5-fold CV)
Method 2b	Logistic Regression	$\beta_0 + \beta_1 Female + \beta_2 YearsOfExperience + \beta_3 Female \times YearsOfExperience$	0.5556 (30% Test Set)	0.5794 (30% Test Set)
	Adaboost	$All\ Features$	0.5132 (5-fold CV)	0.5235 (5-fold CV)
	Decision Trees	$EducSystem + Age + ATS_{Weighted} + EducSystem \times ATS_{Weighted} + ATS_{Weighted}^2$	0.5393 (5-fold CV)	0.5198 (5-fold CV)

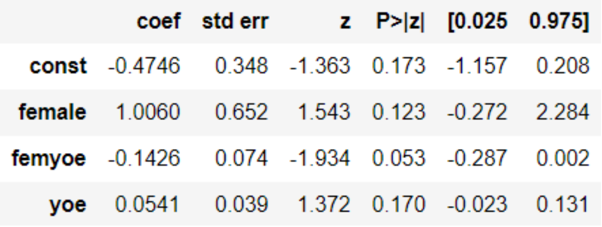
In both methods, Logistic Regression outperforms the other models. The Perceptron was not included in Method 2b as it was concluded to perform worse than random guessing, with accuracy less than 0.5. A more detailed output for Logistic Regression is included below. In the output for method 2a, all predictors are significant for alpha = 0.10 (while the interaction of education system and weighted ATS is not strictly under 0.10, we accept it being significant using 2 decimal places). While the linear terms Female and Years of Experience are insignificant using the same alpha level in Method 2b, their interaction term is significant, thus we must leave the two linear terms in the model.

Table 16. Logistic Regression results from method 2a



	coef	std err	z	P>|z|	[0.025	0.975]
const	-8.9648	3.571	-2.510	0.012	-15.964	-1.966
educsystem	1.3944	0.732	1.905	0.057	-0.040	2.829
age	0.0518	0.028	1.883	0.060	-0.002	0.106
ats_weighted	17.6910	8.169	2.166	0.030	1.680	33.702
educsystem*ats_weighted	-2.2606	1.387	-1.629	0.103	-4.980	0.459
ats_weighted_2	-9.9893	5.327	-1.875	0.061	-20.430	0.451

Table 17. Logistic Regression results from method 2b



	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.4746	0.348	-1.363	0.173	-1.157	0.208
female	1.0060	0.652	1.543	0.123	-0.272	2.284
femyoe	-0.1426	0.074	-1.934	0.053	-0.287	0.002
yoe	0.0541	0.039	1.372	0.170	-0.023	0.131

In Method 1, an increase in Age directly translates into an increase in the probability of being classified as "Self-aware", ceteris paribus. Because it is a linear term, a unit increase in Age results in a 0.05 increase in the log odds of the ratio of success to failure (log(P/(1-P)), where P is the probability of being self-aware). This is intuitively plausible since as people get older, they have more work experience from which they can draw from. The importance of a candidate's educational background and development is also found in this method, where it plays a non-linear role. Literature has shown that "performance, self-evaluation, and self-representation are systematically interrelated" and "educated persons performed better and rated themselves accordingly across all domains" (Demetriou, 2007).

The variable "Female" can be interpreted slightly differently in Method 2b, as it is part of an interaction term. Should "Female" equal 1, the summation of terms multiplied by their coefficients (before it is passed to the sigmoid function) will increase by 1.0060-(0.1426*YearsOfExp). Thus should a candidate have less than 7 years of experience, their probability of being self-aware increases if they are female. On the other hand, should a candidate have more than 7 years of experience, their probability of being self-aware will increase if they are male. This is in line with our gender self-awareness hypothesis testing results, where we determined that the mean self-awareness for males is different to that of females within the N=275 dataset, and is further backed by research, where it is said that "gender differences do exist, both in rated self-awareness and in one of its subcomponents, knowledge of self" (Velsor, Taylor, and Leslie, 1993)

On the other hand, a unit increase in Years of Experience will see an increase in the summation of terms multiplied by their coefficients (before it is passed to the sigmoid function) by 0.0541-(0.1426*Female). Thus, if a candidate is female, a unit increase in Years of Experience will have a negative impact on their probability of being self-aware, and if they are male, it will increase their log-odds of being self-aware.

Further improvements can be made to the testing and modelling of Methods 2a and 2b through use of cross-validation on the logistic regression models. Currently, due to time constraints, a manual implementation of cross-validation on the logistic regression package in "statsmodels.api" could not be completed. The reason for this is due to SkLearn not offering statistical analyses on their models. A wider search and test on other interaction terms could also be attempted in the future, as we were limited due to time constraints.

5.3.3 Method Comparison and Final Specification

From the two methods of classifying candidates' self-awareness, by comparing their predictive performances, method 2a achieved a slightly better ROC_AUC score compared to the score produced by method 1. However, method 2 only works if there exists data on previous candidates who applied for the job and the accuracy of the model wouldn't be valid if the employer has decided to modify the test. Hence, we select method 1 as our final classification model because it aligns with existing economic literature and the only requirement for this method is for Alooba to justify a range of the self-awareness score for "self-aware" candidates to fall into based on the difficulty of the specific test they took. The only limitation for method 1 is that it produces a very high imbalanced class. We suggest that

Alooba should look into methods of dealing with imbalanced classes such as over-sampling or under-sampling to improve the overall predictive model and reduce bias on the majority class.

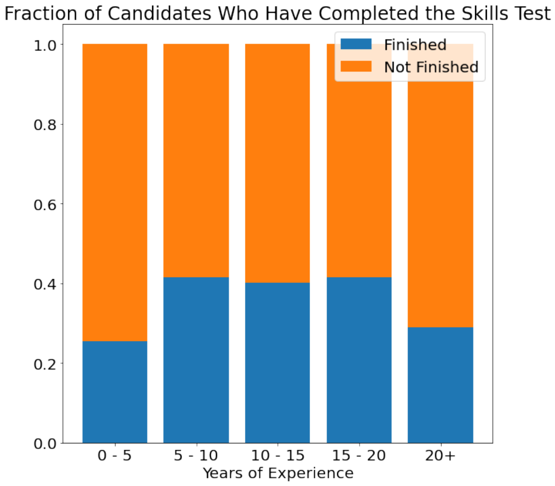
5.4 Additional Findings

5.4.1 Years of Experience and Test Completion

We found an interesting relationship between a candidate's prior experience and their test taking behaviours in our initial hypothesis testing. We identified that experienced candidates and recent graduates both have lower rates of test completion.

Our null hypothesis that H0: Ratio of candidates completing tests in each years of experience group is the same. If the ratio of people completing the test is roughly equal in each group, then we can say that a candidate's years of experience has no effect on whether or not they will attempt the skills test. A Chi-square test for independence was run, producing the conclusion that the likelihood of a candidate completing the skills test depends on their years of experience. This is shown in the graph below.

Figure 5: Fraction of candidates who have completed the skills test

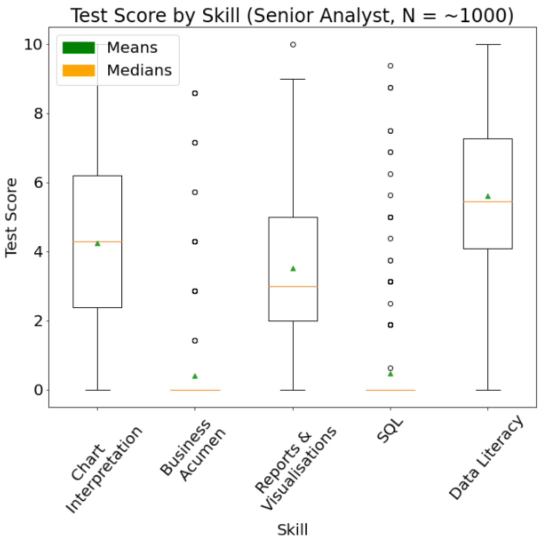


The low percentage of very experienced candidates who complete the skills test (20+ years) has cause for concern, more so than the inexperienced (0-5 years). Past research has shown that skills tests are "valid predictors of successful performance for all jobs in all settings" (Scepura, 2020) thus having more highly experienced candidates with a test score from which we can judge their performance and self-awareness will yield higher quality candidates for a role. The low engagement rate could potentially arise from accessibility issues. Older generations may have more trouble assessing online testing and age discrimination is prevalent in the workplace and job market so to reduce barriers to entry and implicit bias, Alooba should investigate the cause of these low engagement rates and improve user experience (Macdonald and Levy, 2016).

5.4.1 Years of Experience and Test Completion

Within the same job test, we found discrepancies in the mean performance of candidates between questions about different skills. By computing the difference between self-rating for a given skill and the actual skill test performance, we find skill tests where a candidate's actual performance is drastically different from their self-rating score. This could help identify if there are issues in how the tests have been set up which could induce bias within the recruitment process. Using the Friedman test, we conclude that there is at least one skill whose distribution of their "expected" - "actual" metric is different to the others, as shown below.

Figure 6: Test scores by skill for Senior Analyst position



In figure 6 above we see an average mark of 0 in business acumen and SQL tests. Conducting a post-hoc Nemenyi test, we verify that candidates for the Senior Analyst role struggle with estimating their Business Acumen and SQL performance. Both the mean in green and median in yellow are significantly lower than other skills. This is contrary to Data Literacy, in which candidates are closest to estimating their true test score. Business acumen is a qualitative skill and can be difficult to assess. Hence, it is usually assessed during the interview stage of job recruitment (Kaplan, 2006). Alooba should consider examining the validity of including this skill test in their online assessment. Alooba should consider standardizing the difficulty of tests or re-evaluate the styles of assessment to more accurately capture candidate skills.

6. Conclusion

Our results have demonstrated that there is value for Alooba in considering self-awareness of candidates alongside technical skills testing when assessing job candidate fit. We have established a quantitative definition for self-awareness against which we identify different self-awareness between demographic groups and establish a parsimonious classification model for predicting self-aware candidates. We have also identified additional results that have business implications for Alooba and provide recommendations for future product implementation and research.

From textual analysis of candidate CVs, we identify that information contained within a candidate's CV provides explanatory power in assessing candidate fit both by providing predictive significance to test performance and as a predictor of self-awareness. Thus, Alooba should consider replacing manual CV screening with automated processes that adequately capture CV relevance through BOW and ATS ratings. The university a candidate attends is also relevant, so we suggest that Alooba collects this information about clients either in the pre-test questionnaire or automated collection through CV screening NLP. Due to inconsistent CV formatting, we suggest including it in the pre-test questionnaire. Alooba should also consider expanding the data fields in either the job application or pre-test that can be used to fine-tune the self-awareness measure and also track diversity quotas to ensure that their products allow all applicants to fairly partake and do not induce unwarranted bias. Future fields include a non-binary gender option or a cover letter from which NLP methods could extract more succinct CV/application relevance from. We believe that our project has unearthed invaluable insights to Alooba for use in informing development of their systematized and unbiased recruitment products.

References

Andreas Demetriou (2007). *Reasoning and self-awareness from adolescence to middle age: Organization and development as a function of education. From:* https://www.google.com/url?q=https://www.sciencedirect.com/science/article/abs/pii/S104160800800109X&sa=D&source=docs&ust=1637483837971000&usg=AOvVaw2Hdf2Dsx6zQrS_2gsQAGbZ

Balart, P., Oosterveen (2019). M. *Females show more sustained performance during test-taking than males*. Nat Commun 10, 3798. From: https://doi.org/10.1038/s41467-019-11691-y

Bock, T., (2021). *Decision Trees Are Usually Better Than Logistic Regression* https://www.displayr.com/decision-trees-are-usually-better-than-logistic-regression/

Bright, J. , Hutton, S., (2002). *The Impact of Competency Statements on Résumés for Short-listing Decisions* https://doi.org/10.1111/1468-2389.00132

Brownlee, J., (2020). *How to Calculate Feature Importance With Python* https://www.google.com/url?q=https://machinelearningmastery.com/calculate-feature-importance-with-python/&sa=D&source=docs&ust=1637478952225000&usg=AOvVaw1_8Vauadj7pvADDqQmn2lT

Brownlee, J., (2020). *Probabilistic Model Selection with AIC, BIC, and MDL* https://www.google.com/url?q=https://machinelearningmastery.com/probabilistic-model-selection-measures/&sa=D&source=docs&ust=1637483238261000&usg=AOvVaw0PLZyFGMoC9qWCWlAaCZQV

Can Pre-Employment Tests Predict Employee Success Better than a Human?. Criteria Corp. (2021). https://www.criteriacorp.com/can-pre-employment-tests-predict-employee-success-better-human.

Charlton, E (2018). *What living abroad does to your self-awareness,* World Economic Forum. https://www.weforum.org/agenda/2018/06/considering-an-overseas-move-living-abroad-can-boost-your-self-awareness-and-help-your-career/

Dodd-McCue, D., & Tartaglia, A. (2010). *Self-report Response Bias: Learning How to Live with its Diagnosis in Chaplaincy Research.* Chaplaincy Today, 26(1), 2-8. https://doi.org/10.1080/10999183.2010.10767394

Fletcher, C., Bailey, C., (2003). *Assessing self-awareness: some issues and methods*. Journal of Managerial Psychology, 18(5), 395–404. https://www.emerald.com/insight/content/doi/10.1108/02683940310484008/full/html

Glen, S. (2021). *"Correlation Coefficient: Simple Definition, Formula, Easy Steps" From StatisticsHowTo.com: Elementary Statistics for the rest of us!* https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/

Gusdorf, P (2018). *Recruitment and Selection: Hiring the Right Person* https://www.shrm.org/certification/educators/Documents/Recruitment%20and%20Selection%20IM.pdf

Hansford, B. C., & Hattie, J. A. (1982). *The Relationship between Self and Achievement/Performance Measures.* Review of Educational Research, 52(1), 123–142. https://doi.org/10.2307/1170275

Jayawant N.Mandrekar (2015). *Receiver Operating Characteristic Curve in Diagnostic Test Assessment From:*

https://www.google.com/url?q=https://www.sciencedirect.com/science/article/pii/S1556086415306043&sa=D&source=docs&ust=1637478994690000&usg=AOvVaw2MFof_Pu46jmZYgewO1yAQ

Knight, R. (2017). *7 Practical Ways to Reduce Bias in Your Hiring Process*. Harvard Business Review. https://hbr.org/2017/06/7-practical-ways-to-reduce-bias-in-your-hiring-process

Kudroli Foundation (2019). *The Importance of Education in Developing Countries* https://www.google.com/url?q=https://www.kudroli.org/blogs/the-importance-of-education-in-developing-countries&sa=D&source=docs&ust=1637482848256000&usg=AOvVaw0MXEYqIlAOHh6F9j6fJ5zt

Laumer, S., Maier, C. & Eckhardt, A. (2015). *The impact of business process management and applicant tracking systems on recruiting process performance: an empirical study*. Journal of Business Economics, vol 85, 421–453. https://doi.org/10.1007/s11573-014-0758-9

Macdonald, J., Levy, S.R. (2016). *Ageism in the Workplace: The Role of Psychosocial Factors in Predicting Job Satisfaction, Commitment, and Engagement* From: *https://www.google.com/url?q=https://cpb-us-e1.wpmucdn.com/you.stonybrook.edu/dist/e/2677/files/2018/04/Macdonald_Levy_2016pdf-1m7pfez.pdf&sa=D&source=docs&ust=1637491794246000&usg=AOvVaw2o4PaVCFnPp7wMgahnrBpe*

McLean, VA (2019). *New Index of Excellence in STEM Education Compares U.S. Students to Global Competition* https://www.google.com/url?q=https://www.cee.org/news-events/events/new-index-excellence-stem-education&sa=D&source=docs&ust=1637481630810000&usg=AOvVaw1q2qvIsMSiRNE7LWKQYtFv

Miksch, D. (2018). Countering the Consequences of Objective Self-awareness [Unpublished Masters thesis]. Northern Illinois University.

Mujtaba, H. (2020). *An Introduction to Bag of Words (BoW) | What is Bag of Words?* https://www.mygreatlearning.com/blog/bag-of-words/

Nisbett, Richard E; Wilson, Timothy D (1977). *The halo effect: Evidence for unconscious alteration of judgments.* Journal of Personality and Social Psychology. American Psychological Association. 35 (4): 250–56. doi:10.1037/0022-3514.35.4.250. hdl:2027.42/92158.

Scepura, R.,(2020). *The Challenges With Pre-Employment Testing and Potential Hiring Bias* https://doi.org/10.1016/j.mnl.2019.11.014

Shields, J. (2018). *Taleo: 4 Ways the Most Popular ATS Rates Your Resume* https://www.jobscan.co/blog/taleo-popular-ats-ranks-job-applications/.

Shurchkov, O. (2012). *Under pressure: gender differences in output quality and quantity under competition and time constraints.* From: https://www.google.com/url?q=https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1542-4774.2012.01084.x&sa=D&source=docs&ust=1637477066391000&usg=AOvVaw0iPrejgnihPBRNIk9WqlNA

Strietholt, R., Bos, W., Gustafsson, J., Rosén, M., (2014). *Educational Policy Evaluation through International Comparative Assessments.* Published by Cengage, London.

Tsai, Wei-Chi et al. *The Effects Of Applicant Résumé Contents On Recruiters' Hiring Recommendations: The Mediating Roles Of Recruiter Fit Perceptions. Applied Psychology*, vol 60, no. 2, 2010, pp. 231-254. Wiley, https://doi.org/10.1111/j.1464-0597.2010.00434.x

Van Velsor, E., Taylor, S., Leslie, J. (1993). *An examination of the relationships among self-perception accuracy, self-awareness, gender, and leader effectiveness* https://www.google.com/url?q=https://doi.org/10.1002/hrm.3930320205&sa=D&source=docs&ust=1637481960104000&usg=AOvVaw12HEtujSWulndItAh9sbau

What Self-Awareness Really Is (And How To Cultivate It). Harvard Business Review (2021), https://hbr.org/2018/01/what-self-awareness-really-is-and-how-to-cultivate-it

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, *48*(4), 817–838. https://doi.org/10.2307/1912934

Will Kenton (2021). *Analysis of Variance (ANOVA)* https://www.investopedia.com/terms/a/anova.asp

Wylie, R. (2019). *The self-concept: Theory and research on selected topics (Vol. 2)*. Lincoln: University of Nebraska Press

Appendix

Appendix 2.1- Peer Review Feedback

Here we present the peer review we received from our draft and detail how we addressed the suggestions.

Feedback 1:

This draft report is so well-presented and informative. Deriving self-awareness and achieving a distribution close to normal distribution is excellent. Also, all results and plots are sufficient and directly link to the objectives that the team wants to achieve.

There are two things that our group found can be improved on:

1. When doing the OLS analysis, the group used most, if not all variables, to predict the overall test score. The model has a low R-square, data transformations will be beneficial. The report didn't mention the rationale of choosing a quadratic form of certain variables, like age. There exist parameters, whose p-value is > 0.05. Although the model did mention them, it would be great to know how to deal with this problem, i.e. removing the? ***We addressed this suggestion by clearly detailing our feature engineering and variable selection for the OLS model. Refer to section 4.2 in the report.***

2. The report seems to spend more focus on the EDA rather than the prediction model of the data. The prediction model should be done after the hypothesis is induced from the EDA. Also, the efficiency of comparing the self awareness with all independent variables is doubted. ***There was a greater focus on EDA in the draft as the final model output was not ready. It has been detailed in section 5 of our report. We reorganized the report to separate out empirical model setup from results. We provide robust testing to justify the efficiency of comparing self-awareness with the subset of variables used in our final results.***

Feedback 2:

Positives

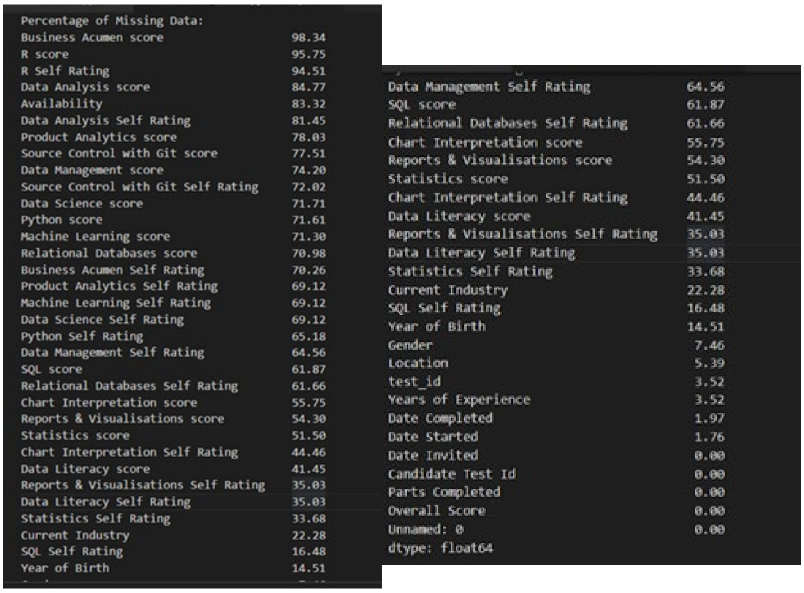
A very detailed and thorough background research provided an excellent starting point and understanding which allowed the reader to approach the report in an informed manner. Furthermore, the methodology and overall structure of the report was clear and homogenous, enabling a comprehensive reading. ***Thank.***

Improvements

Our major critique involves the lack of explanation upon how the interaction terms for the linear regression were selected. Some interaction terms like gender/age and gender/self-awareness rating were included, however, many other interaction terms such as the interaction between gender and country have been omitted. ***Detailed justification about correlation testing to determine interaction terms has been included in the final report following this feedback. Refer to section 4.2.*** Different countries have varying levels of gender equality, and therefore, the relationship between men and women is treated differently depending on the country. The reasoning for selecting some interaction terms and not others should be mentioned. ***As above, we considered this by including justification.*** Secondly, a minor critique, if India and the UAE are being treated as their own dummy variables - due to their large sample size - then they should be excluded from the Asia dummy variable, or a justification would be necessary. ***We have justified our decision to still include UAE and India in the dummy variable for Asia in section 3.1.***

Appendix 3.1- Missing Values

Figure 7: Missing values from combined_results.xlsx



```
Percentage of Missing Data:
Business Acumen score                   98.34
R score                                 95.75
R Self Rating                           94.51
Data Analysis score                     84.77
Availability                            83.32
Data Analysis Self Rating               81.45
Product Analytics score                 78.03
Source Control with Git score           77.51
Data Management score                   74.20
Source Control with Git Self Rating     72.02
Data Science score                      71.71
Python score                            71.61
Machine Learning score                  71.30
Relational Databases score              70.98
Business Acumen Self Rating             70.26
Product Analytics Self Rating           69.12
Machine Learning Self Rating            69.12
Data Science Self Rating                69.12
Python Self Rating                      65.18
Data Management Self Rating             64.56
SQL score                               61.87
Relational Databases Self Rating        61.66
Chart Interpretation score              55.75
Reports & Visualisations score          54.30
Statistics score                        51.50
Chart Interpretation Self Rating        44.46
Data Literacy score                     41.45
Reports & Visualisations Self Rating    35.03
Data Literacy Self Rating               35.03
Statistics Self Rating                  33.68
Current Industry                        22.28
SQL Self Rating                         16.48
Year of Birth                           14.51
```

```
Data Management Self Rating             64.56
SQL score                               61.87
Relational Databases Self Rating        61.66
Chart Interpretation score              55.75
Reports & Visualisations score          54.30
Statistics score                        51.50
Chart Interpretation Self Rating        44.46
Data Literacy score                     41.45
Reports & Visualisations Self Rating    35.03
Data Literacy Self Rating               35.03
Statistics Self Rating                  33.68
Current Industry                        22.28
SQL Self Rating                         16.48
Year of Birth                           14.51
Gender                                   7.46
Location                                 5.39
test_id                                  3.52
Years of Experience                      3.52
Date Completed                           1.97
Date Started                             1.76
Date Invited                             0.00
Candidate Test Id                        0.00
Parts Completed                          0.00
Overall Score                            0.00
Unnamed: 0                               0.00
dtype: float64
```

Figure 8: Missing values from senior_commerial_analyst.xlsx (n=965)

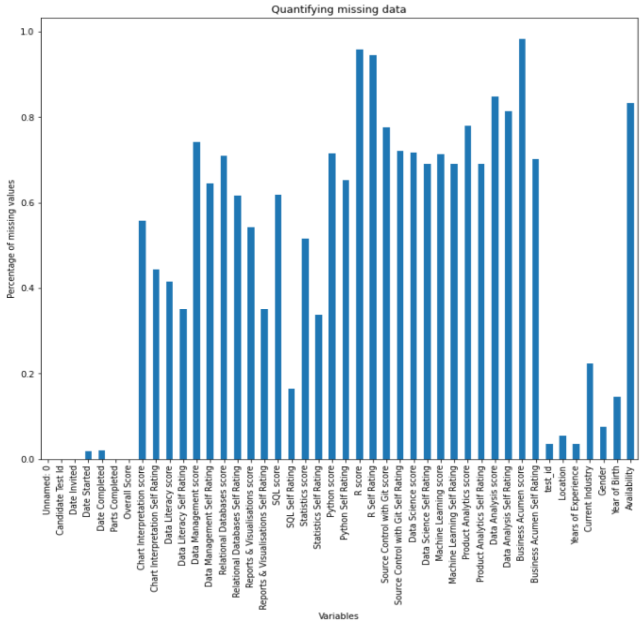
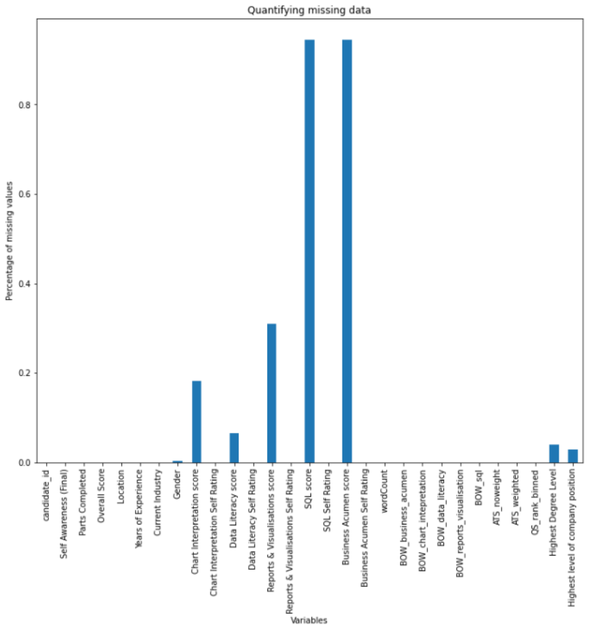


Figure 9: A subset of combined_results dataset (senior commercial analyst) (n=275)



Based on the combined_results dataset, we can see that the largest proportion of missing values lies in candidates' Business Acumen score (98.34%), followed by R score (95.75%), R Self Rating (94.51%) and along with other variables as shown in figure 1 (a).

The Senior Commercial Analyst dataset which is a subset of the combined_results dataset has 275 observations. From figure 1 (b), we observe that candidates' Business Acumen score and SQL score both represent the largest proportion of missing values in our dataset (94.55 %) and this is followed by their Reports and Visualizations score (30.91 %) and Chart Interpretation score (18.18). We also note that there are missing values which lie in the Data Literacy score (6.55 %), and other information relating directly to individual candidates (4 % for highest degree level, 2.91 % for highest level of company position, and 0.36 % for gender).

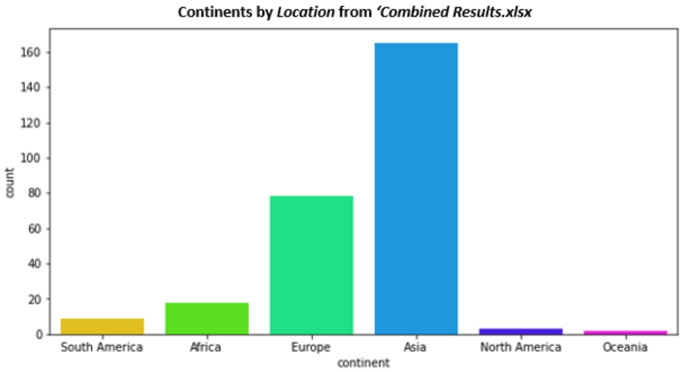
Appendix 3.2

Table 16. Count of candidates in each countries

Countries by *Location* from *'Combined Results.xlsx'*	**Count**
United Arab Emirates	87
India	35
United Kingdom	14
Portugal	14
Ireland	10
Israel	10
Other	104

There were 45 unique countries found in the *Senior_Commercial_Analyst* dataset. In order to simplify our modelling, we categorised each country into the continents they belong to and found that Asia received the most applications, followed by Europe.

Figure 10: Continents by Location from combined_results.xlsx



Appendix 3.3

Building our ATS weighted variable

Current Applicant Tracking Software (ATS) algorithms look for relevant hard skills present in a candidate's resume. These include technical skills, credentials, position titles, relevant software and tools. Soft skills are usually manually assessed via the cover letter or an interview, as commonly used resume keywords and phrases such as "team player" are not quantifiable. These algorithms work by matching exact terms found on the job description, and weighing them in accordance with how important the recruiter deems the skill to be in determining job fit. We take this approach by identifying a set of "must have" skills and "nice have" skills present in the job description.

Table 17. "Must Have" skills and "Nice Have" skills listed in the Senior Commercial Analyst Job Description

"Must Have" Skills	**"Nice Have" Skills**
Sql, excel, looker, data, analysis, analyst, dashboard, data literacy, project management, English, commercial	Business, management, python, r programming, entrepreneurial, travel, consulting, scrum

We counted the occurrence of each of these skills in each CV and set a dummy variable of 1=skill present and 0=skill not present. We then summed up these scores and weighted the "must have" skills by 150% and 100% on the "nice have" skills and divided this score by N=18, the total number of skills to create *ats_weighted.*

We explore different methods of weighting the significance of each set of skills and determined that the distribution of a 150% weight on "must have" skills was best fitted as a variable in predicting overall score compared against other weightings and an unweighted score.

Relevant code provided below in appendix 3.4.1 and in the google drive.

Appendix 3.4

Building our bow_avgN variable

We identify the following skills and characteristics to be of relevance to the skills tested by Alooba, as sourced from various articles on the internet. Assumptions were made about this relevance as we did not have access to test questions given to this set of candidates.

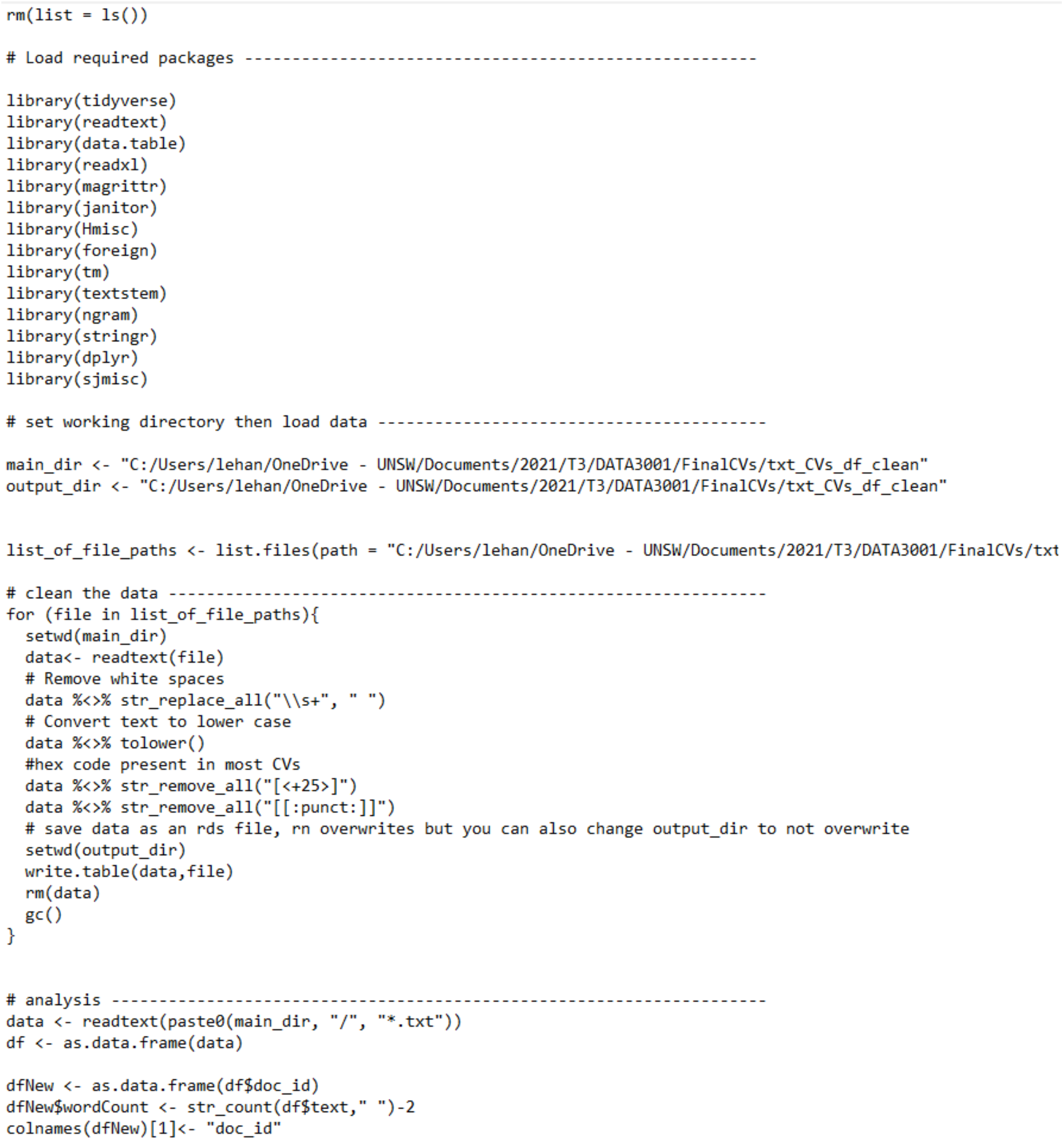
Table 18. Skills tests and their relevant keywords

Skills test	**Relevant words**

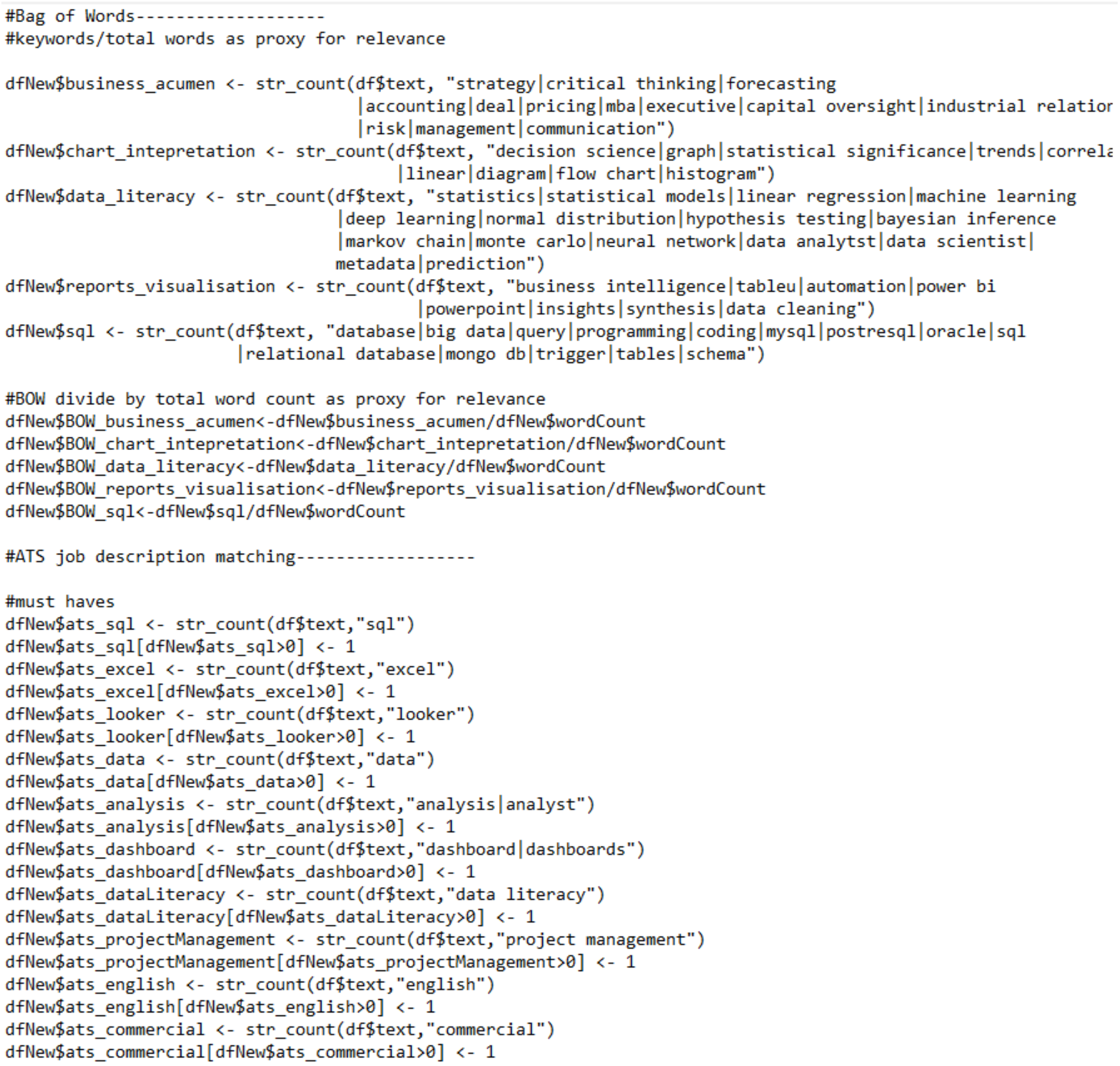
Business Acumen	Strategy, critical thinking, forecasting, accounting, deal, pricing, mba, executive, capital oversight, industrial relations, risk, management, communication
Chart Interpretation	decision science, graph, statistical significance, trends, correlation, linear, diagram, flow chart, histogram
Data literacy	Statistics, statistical models, linear regression, machine learning, deep learning, normal distribution, hypothesis testing, bayesian inference, markov chain, monte carlo, neural network, data analyst[LZ1] , data scientist, metadata, prediction"
Reports and visualisation	business intelligence, tableau, automation, power bi, powerpoint, insights, synthesis, data cleaning
SQL	Database, big data, query, programming, coding, mysql, postresql, oracle, sql, relational database, mongo db, trigger, tables, schema

We divided the occurrence of each skill test (a sum of presence of all relevant words) by the total number of words in each CV for individual bow measures and then averaged those scores for an overall bow variable. We then standardized this variable for consistency.

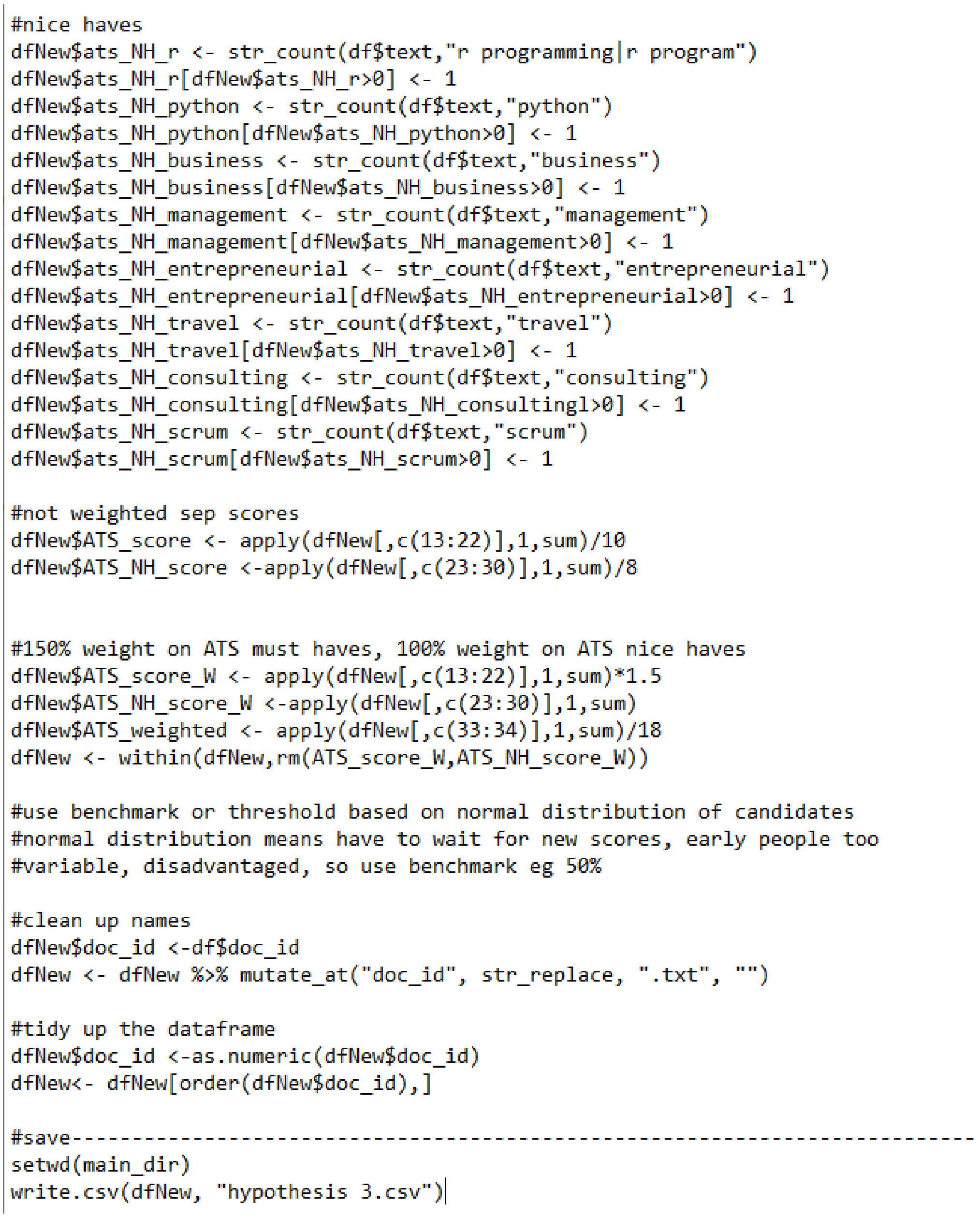
Code for creating ATS_weighted and BOW_weighted variables below.



```
rm(list = ls())

# Load required packages -----------------------------------------------------

library(tidyverse)
library(readtext)
library(data.table)
library(readxl)
library(magrittr)
library(janitor)
library(Hmisc)
library(foreign)
library(tm)
library(textstem)
library(ngram)
library(stringr)
library(dplyr)
library(sjmisc)

# set working directory then load data ---------------------------------------

main_dir <- "C:/Users/lehan/OneDrive - UNSW/Documents/2021/T3/DATA3001/FinalCVs/txt_CVs_df_clean"
output_dir <- "C:/Users/lehan/OneDrive - UNSW/Documents/2021/T3/DATA3001/FinalCVs/txt_CVs_df_clean"


list_of_file_paths <- list.files(path = "C:/Users/lehan/OneDrive - UNSW/Documents/2021/T3/DATA3001/FinalCVs/txt

# clean the data -------------------------------------------------------------
for (file in list_of_file_paths){
  setwd(main_dir)
  data<- readtext(file)
  # Remove white spaces
  data %<>% str_replace_all("\\s+", " ")
  # Convert text to lower case
  data %<>% tolower()
  #hex code present in most CVs
  data %<>% str_remove_all("[<+25>]")
  data %<>% str_remove_all("[[:punct:]]")
  # save data as an rds file, rn overwrites but you can also change output_dir to not overwrite
  setwd(output_dir)
  write.table(data,file)
  rm(data)
  gc()
}


# analysis -------------------------------------------------------------------
data <- readtext(paste0(main_dir, "/", "*.txt"))
df <- as.data.frame(data)

dfNew <- as.data.frame(df$doc_id)
dfNew$wordCount <- str_count(df$text," ")-2
colnames(dfNew)[1]<- "doc_id"
```



```
#Bag of Words-------------------
#keywords/total words as proxy for relevance

dfNew$business_acumen <- str_count(df$text, "strategy|critical thinking|forecasting
                                   |accounting|deal|pricing|mba|executive|capital oversight|industrial relatior
                                   |risk|management|communication")
dfNew$chart_intepretation <- str_count(df$text, "decision science|graph|statistical significance|trends|correla
                                       |linear|diagram|flow chart|histogram")
dfNew$data_literacy <- str_count(df$text, "statistics|statistical models|linear regression|machine learning
                                 |deep learning|normal distribution|hypothesis testing|bayesian inference
                                 |markov chain|monte carlo|neural network|data analytst|data scientist|
                                 metadata|prediction")
dfNew$reports_visualisation <- str_count(df$text, "business intelligence|tableu|automation|power bi
                                         |powerpoint|insights|synthesis|data cleaning")
dfNew$sql <- str_count(df$text, "database|big data|query|programming|coding|mysql|postresql|oracle|sql
                       |relational database|mongo db|trigger|tables|schema")

#BOW divide by total word count as proxy for relevance
dfNew$BOW_business_acumen<-dfNew$business_acumen/dfNew$wordCount
dfNew$BOW_chart_intepretation<-dfNew$chart_intepretation/dfNew$wordCount
dfNew$BOW_data_literacy<-dfNew$data_literacy/dfNew$wordCount
dfNew$BOW_reports_visualisation<-dfNew$reports_visualisation/dfNew$wordCount
dfNew$BOW_sql<-dfNew$sql/dfNew$wordCount

#ATS job description matching------------------

#must haves
dfNew$ats_sql <- str_count(df$text,"sql")
dfNew$ats_sql[dfNew$ats_sql>0] <- 1
dfNew$ats_excel <- str_count(df$text,"excel")
dfNew$ats_excel[dfNew$ats_excel>0] <- 1
dfNew$ats_looker <- str_count(df$text,"looker")
dfNew$ats_looker[dfNew$ats_looker>0] <- 1
dfNew$ats_data <- str_count(df$text,"data")
dfNew$ats_data[dfNew$ats_data>0] <- 1
dfNew$ats_analysis <- str_count(df$text,"analysis|analyst")
dfNew$ats_analysis[dfNew$ats_analysis>0] <- 1
dfNew$ats_dashboard <- str_count(df$text,"dashboard|dashboards")
dfNew$ats_dashboard[dfNew$ats_dashboard>0] <- 1
dfNew$ats_dataLiteracy <- str_count(df$text,"data literacy")
dfNew$ats_dataLiteracy[dfNew$ats_dataLiteracy>0] <- 1
dfNew$ats_projectManagement <- str_count(df$text,"project management")
dfNew$ats_projectManagement[dfNew$ats_projectManagement>0] <- 1
dfNew$ats_english <- str_count(df$text,"english")
dfNew$ats_english[dfNew$ats_english>0] <- 1
dfNew$ats_commercial <- str_count(df$text,"commercial")
dfNew$ats_commercial[dfNew$ats_commercial>0] <- 1
```



```
#nice haves
dfNew$ats_NH_r <- str_count(df$text,"r programming|r program")
dfNew$ats_NH_r[dfNew$ats_NH_r>0] <- 1
dfNew$ats_NH_python <- str_count(df$text,"python")
dfNew$ats_NH_python[dfNew$ats_NH_python>0] <- 1
dfNew$ats_NH_business <- str_count(df$text,"business")
dfNew$ats_NH_business[dfNew$ats_NH_business>0] <- 1
dfNew$ats_NH_management <- str_count(df$text,"management")
dfNew$ats_NH_management[dfNew$ats_NH_management>0] <- 1
dfNew$ats_NH_entrepreneurial <- str_count(df$text,"entrepreneurial")
dfNew$ats_NH_entrepreneurial[dfNew$ats_NH_entrepreneurial>0] <- 1
dfNew$ats_NH_travel <- str_count(df$text,"travel")
dfNew$ats_NH_travel[dfNew$ats_NH_travel>0] <- 1
dfNew$ats_NH_consulting <- str_count(df$text,"consulting")
dfNew$ats_NH_consulting[dfNew$ats_NH_consultingl>0] <- 1
dfNew$ats_NH_scrum <- str_count(df$text,"scrum")
dfNew$ats_NH_scrum[dfNew$ats_NH_scrum>0] <- 1

#not weighted sep scores
dfNew$ATS_score <- apply(dfNew[,c(13:22)],1,sum)/10
dfNew$ATS_NH_score <-apply(dfNew[,c(23:30)],1,sum)/8


#150% weight on ATS must haves, 100% weight on ATS nice haves
dfNew$ATS_score_W <- apply(dfNew[,c(13:22)],1,sum)*1.5
dfNew$ATS_NH_score_W <-apply(dfNew[,c(23:30)],1,sum)
dfNew$ATS_weighted <- apply(dfNew[,c(33:34)],1,sum)/18
dfNew <- within(dfNew,rm(ATS_score_W,ATS_NH_score_W))

#use benchmark or threshold based on normal distribution of candidates
#normal distribution means have to wait for new scores, early people too
#variable, disadvantaged, so use benchmark eg 50%

#clean up names
dfNew$doc_id <-df$doc_id
dfNew <- dfNew %>% mutate_at("doc_id", str_replace, ".txt", "")

#tidy up the dataframe
dfNew$doc_id <-as.numeric(dfNew$doc_id)
dfNew<- dfNew[order(dfNew$doc_id),]

#save---------------------------------------------------------------------
setwd(main_dir)
write.csv(dfNew, "hypothesis 3.csv")
```

Appendix 4.1

Figure 10: Distributions of Bag of Words for each skill tested. Figure shows that distributions are positively skewed across all skills.

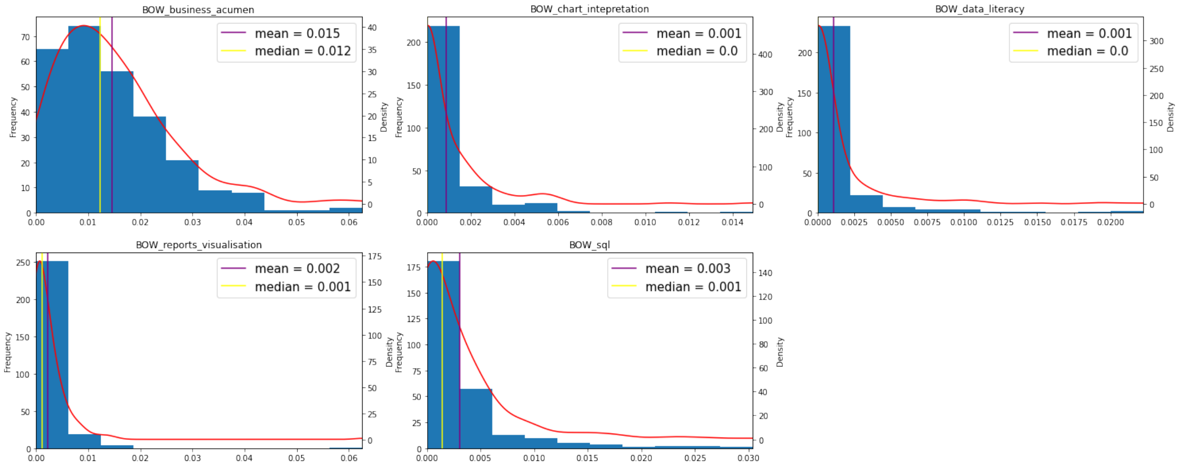


Figure 11: Distribution of ATS weighted score. The distribution is normally distributed.

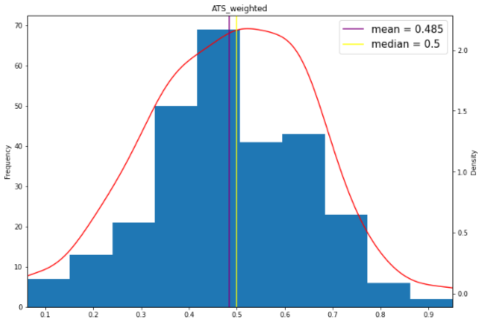
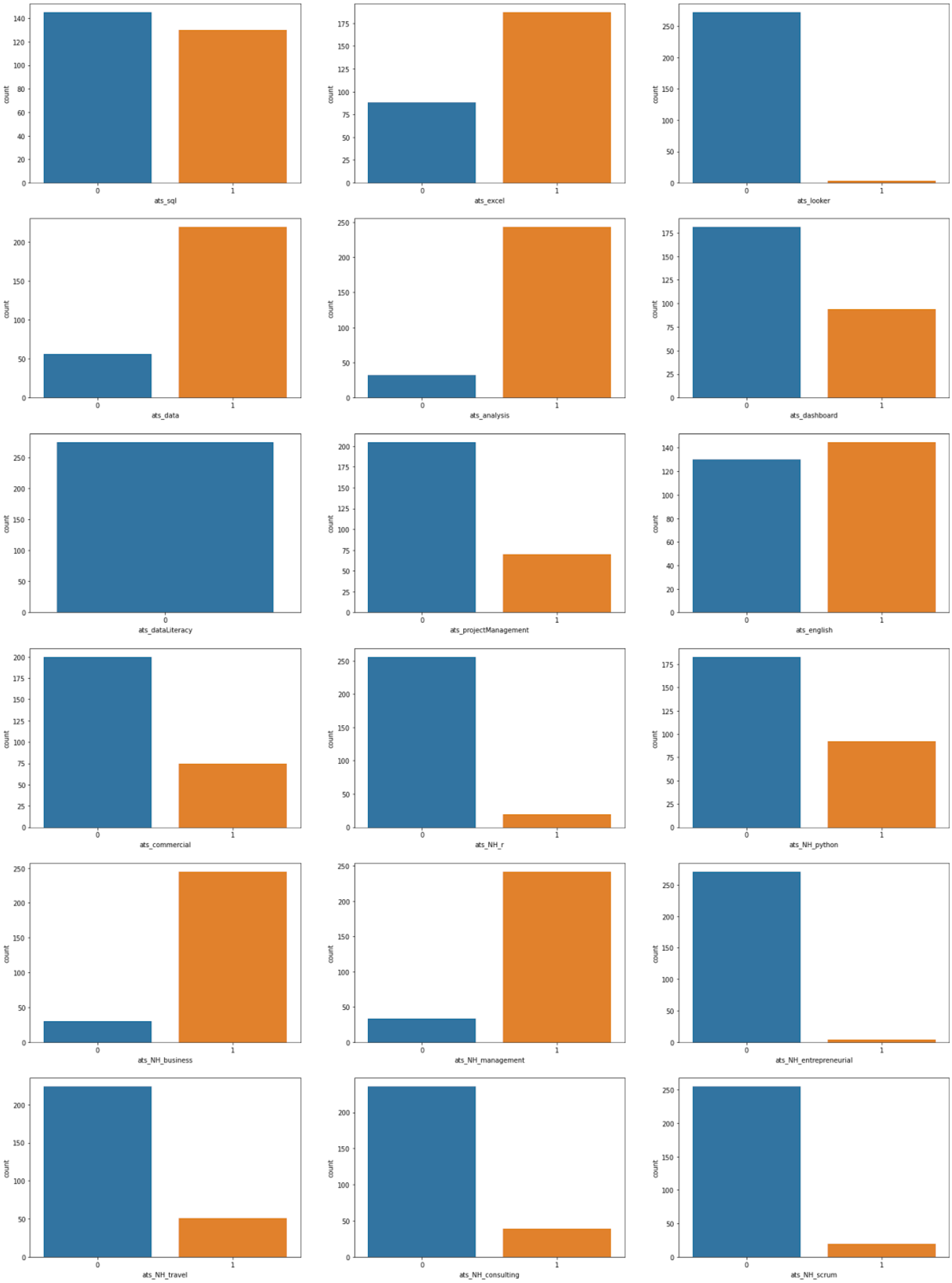
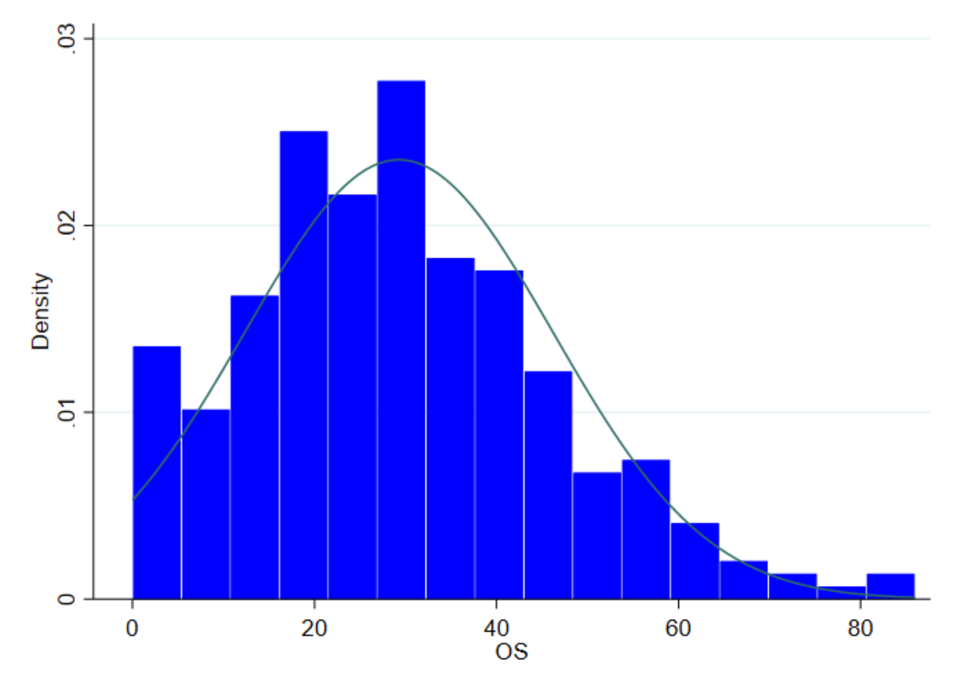


Figure 12: ATS occurrence of keywords founds in job descriptions (must haves and nice haves) extracted from CV. 0 indicates not found and 1 indicates found. Most keywords are not found in majority of candidates except for excel, data analysis and business management


count
0
20
40
60
80
100
120
140
0
1
ats_sql
count
0
25
50
75
100
125
150
175
0
1
ats_excel
count
0
50
100
150
200
250
0
1
ats_looker
count
0
50
100
150
200
0
1
ats_data
count
0
50
100
150
200
250
0
1
ats_analysis
count
0
25
50
75
100
125
150
175
0
1
ats_dashboard
count
0
50
100
150
200
250
0
ats_dataLiteracy
count
0
25
50
75
100
125
150
175
200
0
1
ats_projectManagement
count
0
20
40
60
80
100
120
140
0
1
ats_english
count
0
25
50
75
100
125
150
175
200
0
1
ats_commercial
count
0
50
100
150
200
250
0
1
ats_NH_r
count
0
25
50
75
100
125
150
175
0
1
ats_NH_python
count
0
50
100
150
200
250
0
1
ats_NH_business
count
0
50
100
150
200
250
0
1
ats_NH_management
count
0
50
100
150
200
250
0
1
ats_NH_entrepreneurial
count
0
50
100
150
200
0
1
ats_NH_travel
count
0
50
100
150
200
0
1
ats_NH_consulting
count
0
50
100
150
200
250
0
1
ats_NH_scrum

Appendix 4.2

Figure 13: Distribution for Overall Score. Normally Distributed.



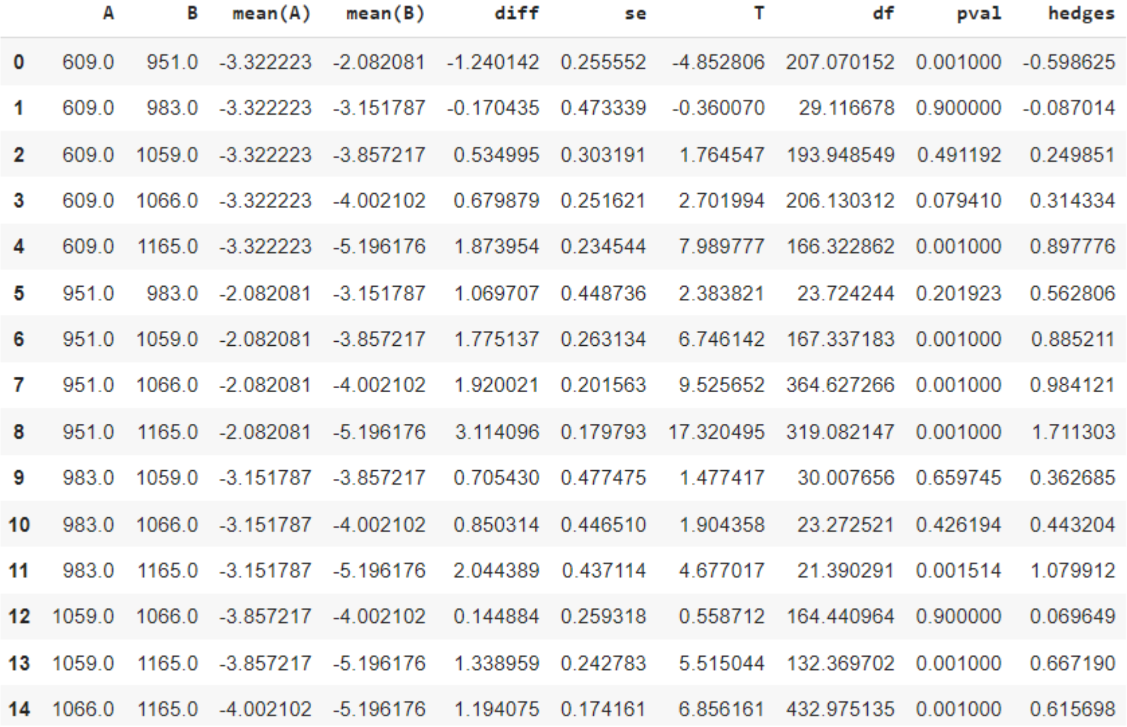
Appendix 4.3

There are altogether 6 distinct test_ids found in the combined_results dataset. We would llike to test if self-awareness scores vary across different Test_ids. With the assumptions that normality holds and that variances are not equal (Barlett's test p-value = 0.029 < 0.05), Welch's Anova test is used to test the following hypothesis:

H0: Means of Self-Awareness are the same across Test_ids vs H1: Means of Self-Awareness are not the same across Test_ids

Under the 5% level of significance, we reject the null hypotheses (p-value = 2.215e-35 < 0.05) and conclude that the means of self-awareness are statistically different across different Test_ids. Games-Howell Test will be used to compare the means of self-awareness and hence identify the different levels of test difficulty. Games-Howell assumes normality, groups of different sizes, variances don't have to be equal, which all hold in each Test_id. From the Games-Howell results, we are able to identify that the mean self awareness for Test_id 1165 is significantly different to all other tests, and it has the highest Self-Awareness Score (negatively), which indicates that the test is more challenging than expected. On the other hand, the mean self awareness for Test_id 951 is significantly different from all other Test_ids except for Test_id 983, and it has the lowest Self-Awareness Score (negatively), which indicates that the test is a lot easier than the other tests, since candidates are able to almost accurately predict their scores.

Table 19. Games-Howell test results - compare means of self-awareness and identify the varying levels of test difficulty

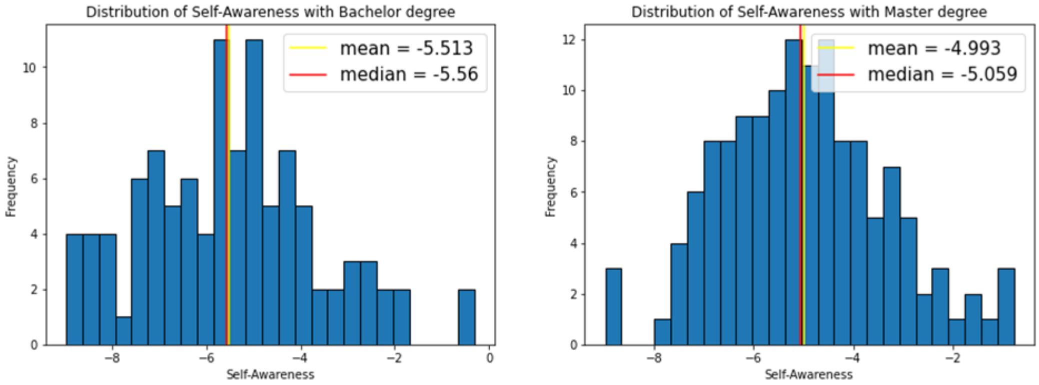


	A	B	mean(A)	mean(B)	diff	se	T	df	pval	hedges
0	609.0	951.0	-3.322223	-2.082081	-1.240142	0.255552	-4.852806	207.070152	0.001000	-0.598625
1	609.0	983.0	-3.322223	-3.151787	-0.170435	0.473339	-0.360070	29.116678	0.900000	-0.087014
2	609.0	1059.0	-3.322223	-3.857217	0.534995	0.303191	1.764547	193.948549	0.491192	0.249851
3	609.0	1066.0	-3.322223	-4.002102	0.679879	0.251621	2.701994	206.130312	0.079410	0.314334
4	609.0	1165.0	-3.322223	-5.196176	1.873954	0.234544	7.989777	166.322862	0.001000	0.897776
5	951.0	983.0	-2.082081	-3.151787	1.069707	0.448736	2.383821	23.724244	0.201923	0.562806
6	951.0	1059.0	-2.082081	-3.857217	1.775137	0.263134	6.746142	167.337183	0.001000	0.885211
7	951.0	1066.0	-2.082081	-4.002102	1.920021	0.201563	9.525652	364.627266	0.001000	0.984121
8	951.0	1165.0	-2.082081	-5.196176	3.114096	0.179793	17.320495	319.082147	0.001000	1.711303
9	983.0	1059.0	-3.151787	-3.857217	0.705430	0.477475	1.477417	30.007656	0.659745	0.362685
10	983.0	1066.0	-3.151787	-4.002102	0.850314	0.446510	1.904358	23.272521	0.426194	0.443204
11	983.0	1165.0	-3.151787	-5.196176	2.044389	0.437114	4.677017	21.390291	0.001514	1.079912
12	1059.0	1066.0	-3.857217	-4.002102	0.144884	0.259318	0.558712	164.440964	0.900000	0.069649
13	1059.0	1165.0	-3.857217	-5.196176	1.338959	0.242783	5.515044	132.369702	0.001000	0.667190
14	1066.0	1165.0	-4.002102	-5.196176	1.194075	0.174161	6.856161	432.975135	0.001000	0.615698

Appendix 5.1

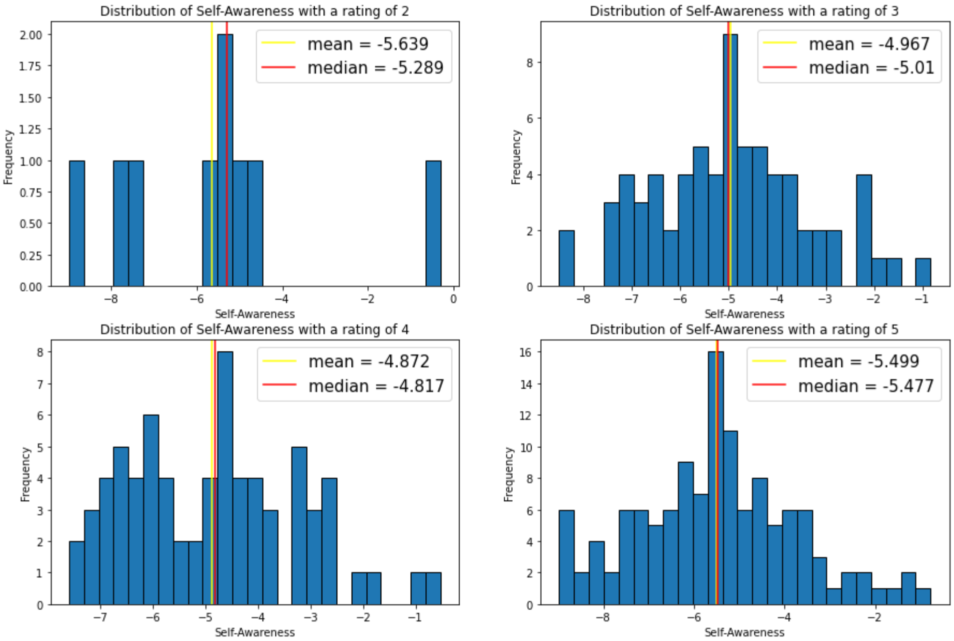
Self-awareness and highest degree level

Figure 14: Distributions of self-awareness for bachelor/master degrees



Self-awareness and Education systems

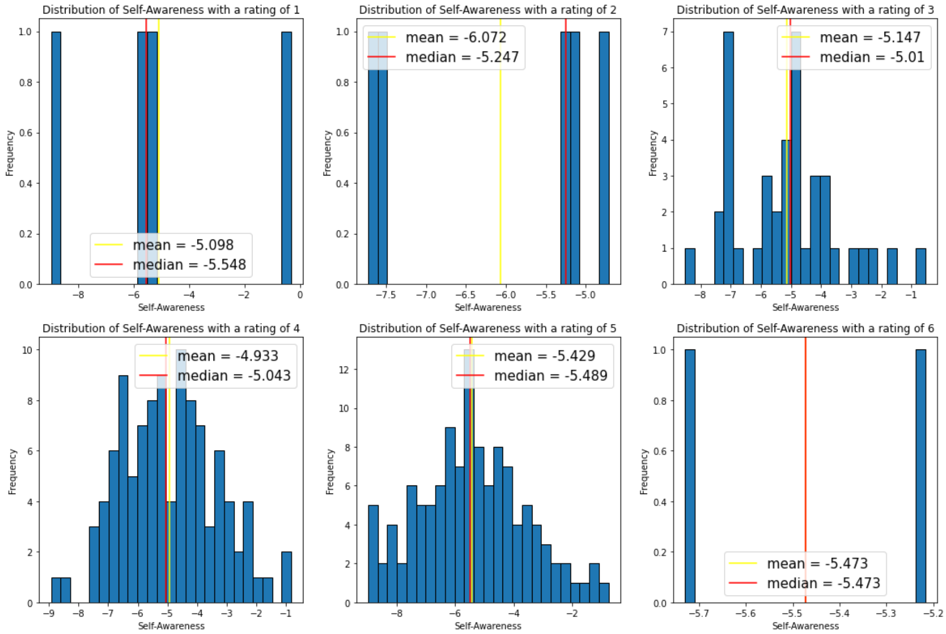
Figure 15. Distributions of self-awareness for different quality of education systems



Self-awareness and maths & science education

Under the 5% level of significance, we accept the null hypothesis (p-value = 0.218 > 0.05) and conclude that the means of self-awareness scores are not statistically different between different quality of maths and science education received by candidates in their countries.

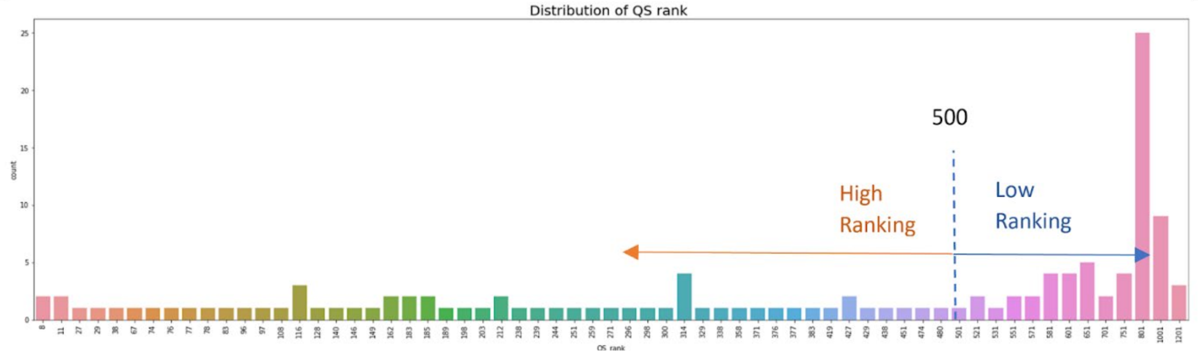
Figure 16. Distributions of self-awareness for different quality of maths & science education



Self-awareness and QS rank

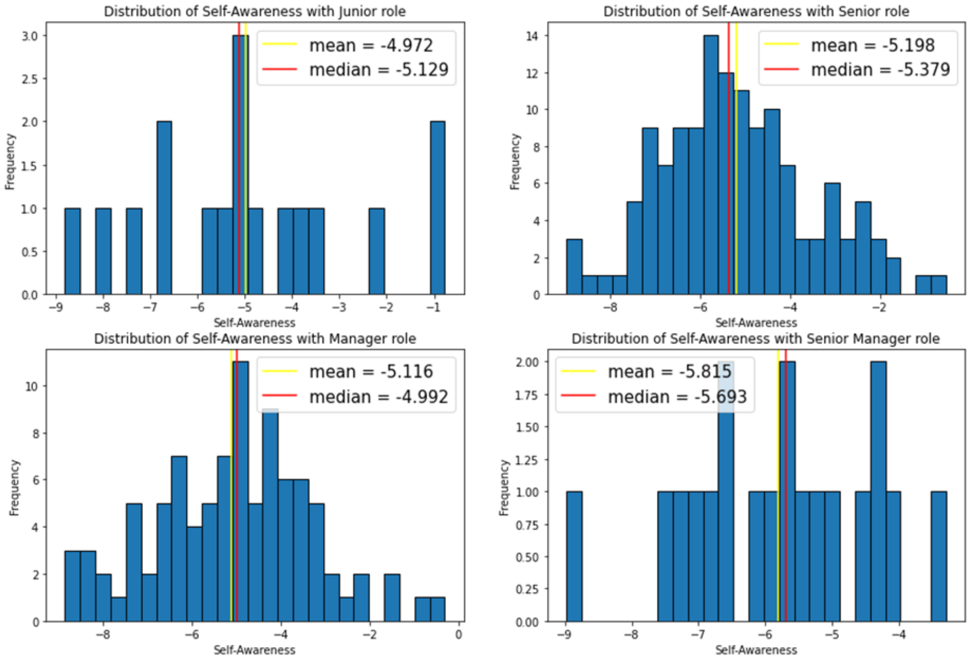
Under the 5% level of significance, we accept the null hypothesis (p-value = 0.915 > 0.05) and conclude that the means of self-awareness score are not statistically different between different QS university rankings.

Figure 17. Distributions of QS rank



Self-awareness and highest company position
Under the 5% level of significance, we accept the null hypothesis (p-value = 0.418 > 0.05) and conclude that means of self-awareness are not statistically different between differing company positions.

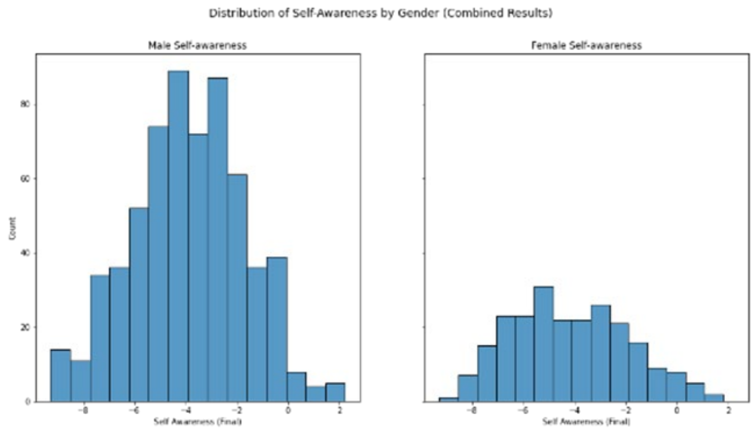
Figure 18. Distributions of self-awareness for different company positions

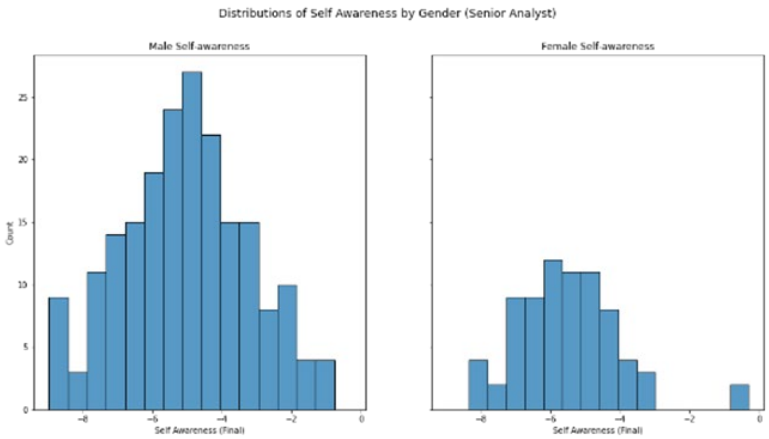


Appendix 5.2

Self-awareness and gender

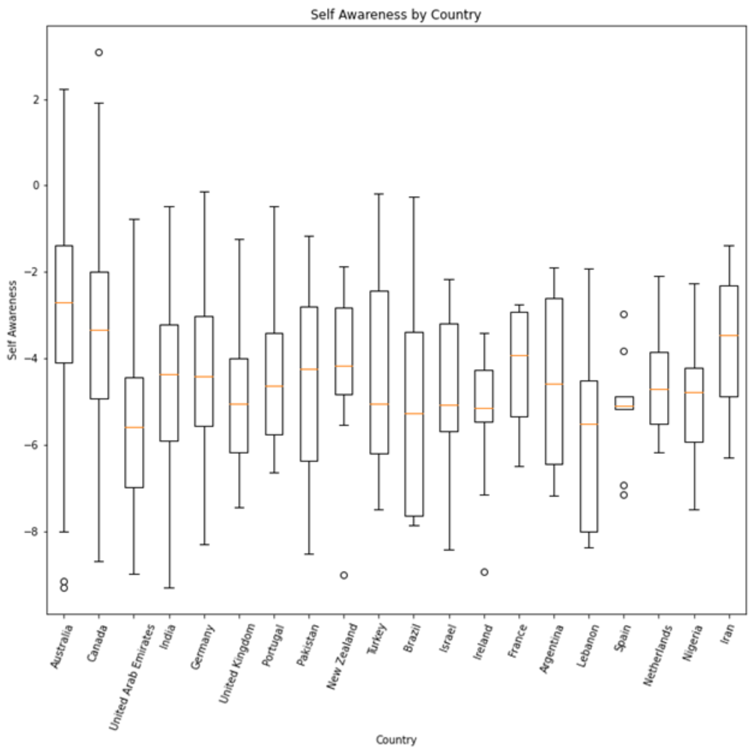
Figure 19. Distributions of self-awareness by Gender (Combined Results & Senior Analyst)

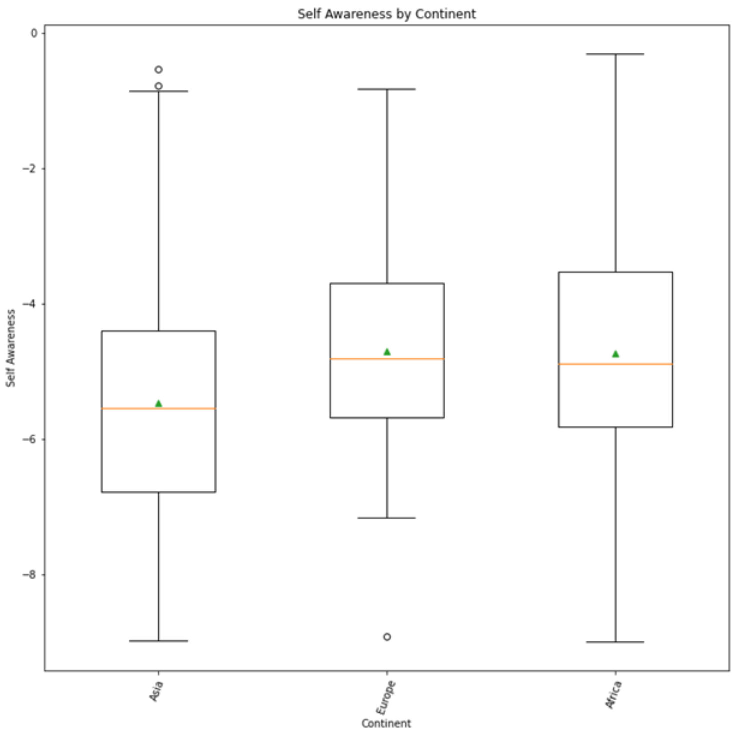




Self-awareness and location of applicant

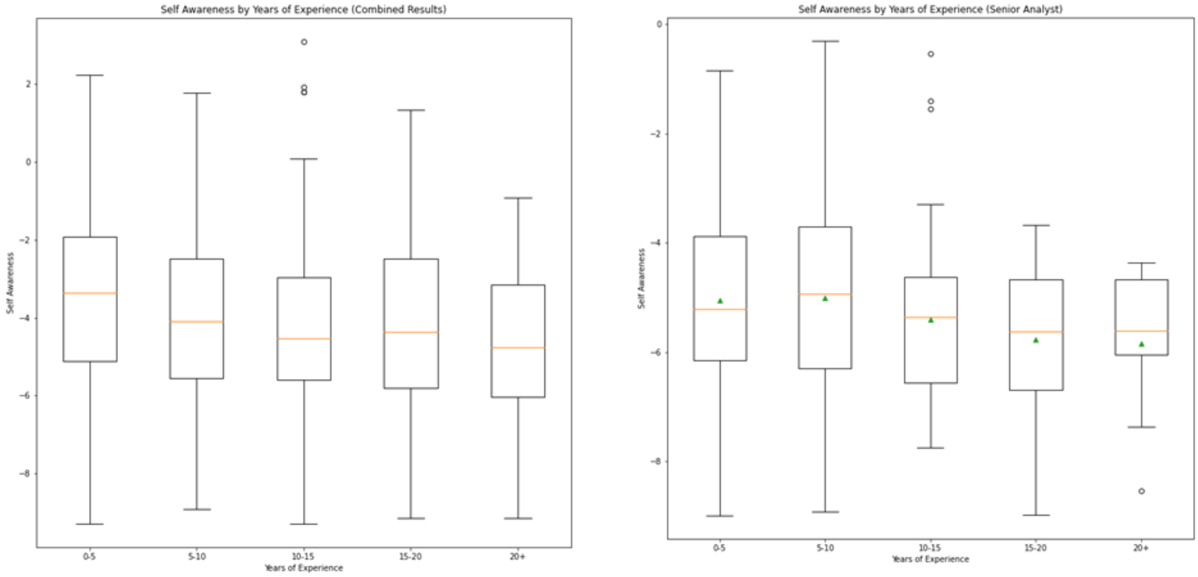
Figure 20. Distributions of self-awareness by Country (top: Combined Results, bottom: Senior Analyst)



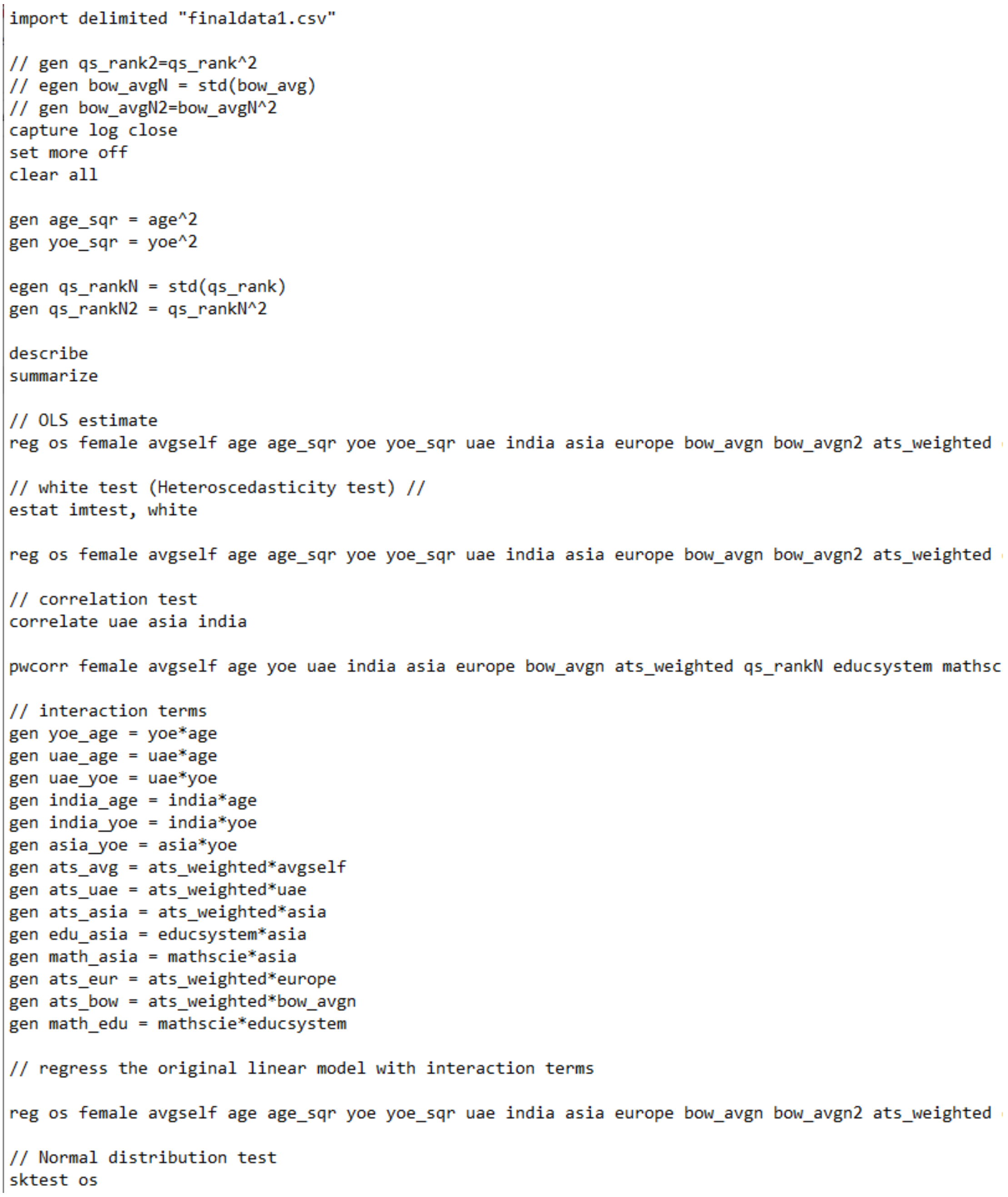


Self-awareness and Years of Experience

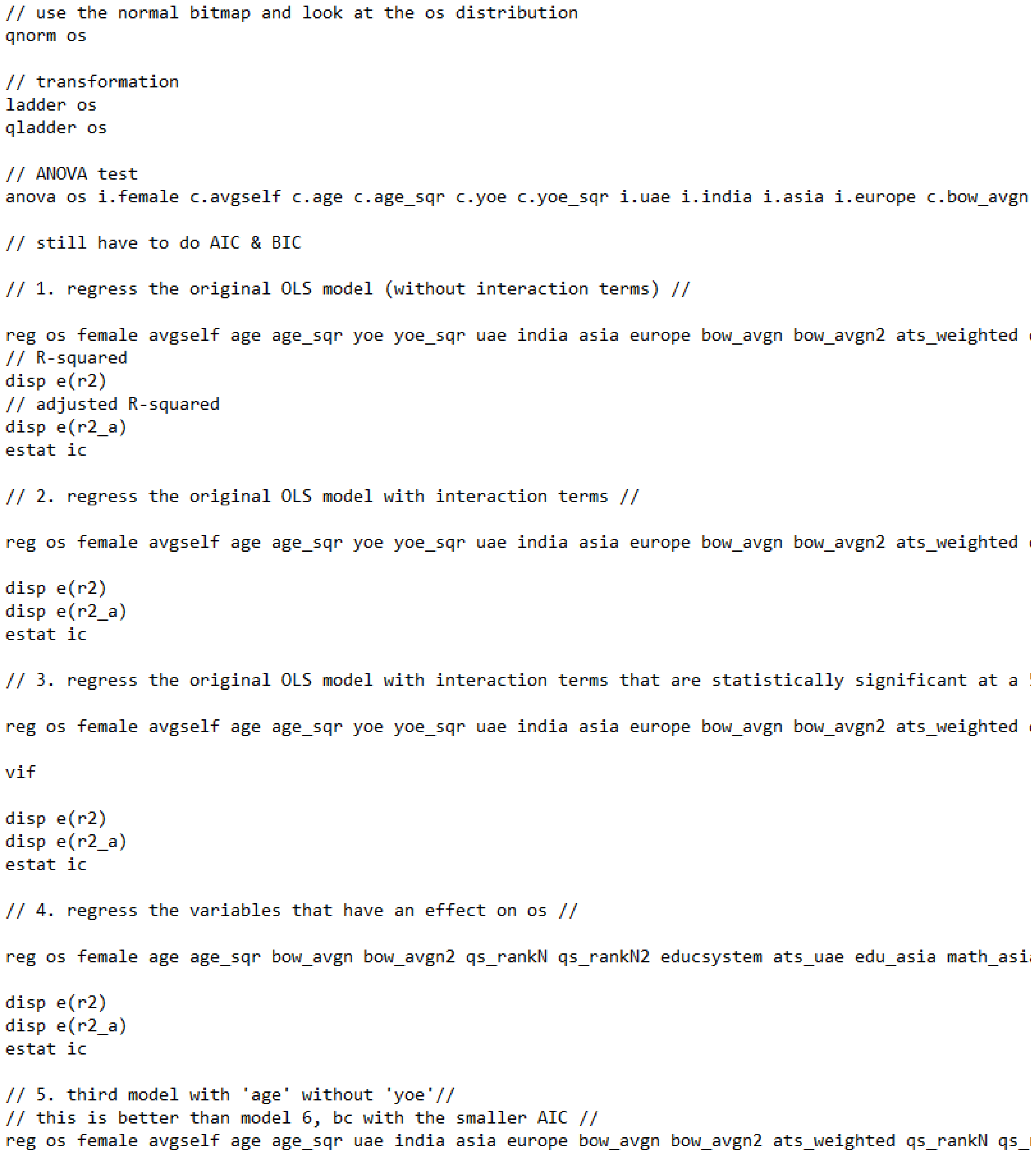
Figure 21. Distributions of self-awareness by Years of Experience (Combined Results & Senior Analyst)



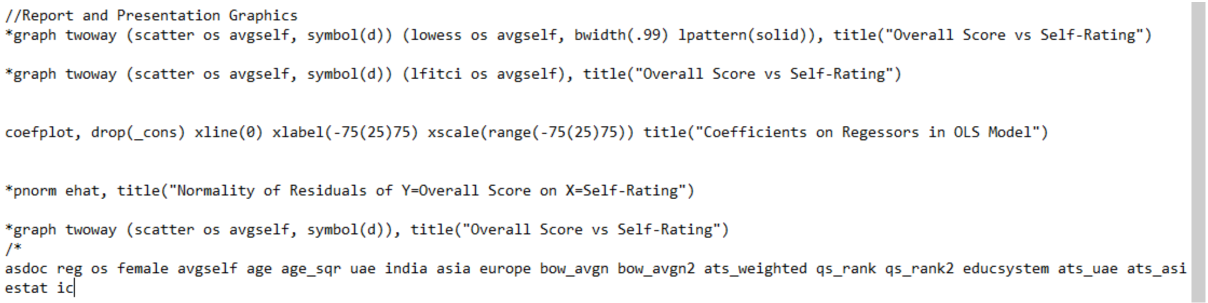
Project Code Appendix



```
import delimited "finaldata1.csv"

// gen qs_rank2=qs_rank^2
// egen bow_avgN = std(bow_avg)
// gen bow_avgN2=bow_avgN^2
capture log close
set more off
clear all

gen age_sqr = age^2
gen yoe_sqr = yoe^2

egen qs_rankN = std(qs_rank)
gen qs_rankN2 = qs_rankN^2

describe
summarize

// OLS estimate
reg os female avgself age age_sqr yoe yoe_sqr uae india asia europe bow_avgn bow_avgn2 ats_weighted

// white test (Heteroscedasticity test) //
estat imtest, white

reg os female avgself age age_sqr yoe yoe_sqr uae india asia europe bow_avgn bow_avgn2 ats_weighted

// correlation test
correlate uae asia india

pwcorr female avgself age yoe uae india asia europe bow_avgn ats_weighted qs_rankN educsystem mathsc

// interaction terms
gen yoe_age = yoe*age
gen uae_age = uae*age
gen uae_yoe = uae*yoe
gen india_age = india*age
gen india_yoe = india*yoe
gen asia_yoe = asia*yoe
gen ats_avg = ats_weighted*avgself
gen ats_uae = ats_weighted*uae
gen ats_asia = ats_weighted*asia
gen edu_asia = educsystem*asia
gen math_asia = mathscie*asia
gen ats_eur = ats_weighted*europe
gen ats_bow = ats_weighted*bow_avgn
gen math_edu = mathscie*educsystem

// regress the original linear model with interaction terms

reg os female avgself age age_sqr yoe yoe_sqr uae india asia europe bow_avgn bow_avgn2 ats_weighted

// Normal distribution test
sktest os
```

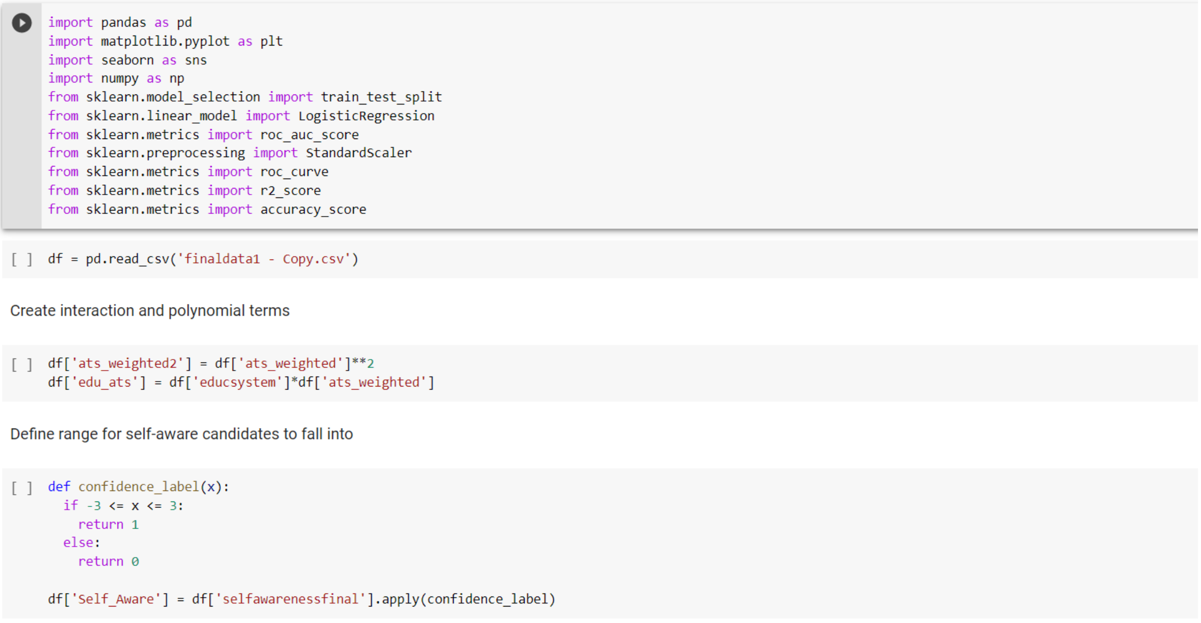


```
// use the normal bitmap and look at the os distribution
qnorm os

// transformation
ladder os
qladder os

// ANOVA test
anova os i.female c.avgself c.age c.age_sqr c.yoe c.yoe_sqr i.uae i.india i.asia i.europe c.bow_avgn

// still have to do AIC & BIC

// 1. regress the original OLS model (without interaction terms) //

reg os female avgself age age_sqr yoe yoe_sqr uae india asia europe bow_avgn bow_avgn2 ats_weighted
// R-squared
disp e(r2)
// adjusted R-squared
disp e(r2_a)
estat ic

// 2. regress the original OLS model with interaction terms //

reg os female avgself age age_sqr yoe yoe_sqr uae india asia europe bow_avgn bow_avgn2 ats_weighted

disp e(r2)
disp e(r2_a)
estat ic

// 3. regress the original OLS model with interaction terms that are statistically significant at a

reg os female avgself age age_sqr yoe yoe_sqr uae india asia europe bow_avgn bow_avgn2 ats_weighted

vif

disp e(r2)
disp e(r2_a)
estat ic

// 4. regress the variables that have an effect on os //

reg os female age age_sqr bow_avgn bow_avgn2 qs_rankN qs_rankN2 educsystem ats_uae edu_asia math_asi

disp e(r2)
disp e(r2_a)
estat ic

// 5. third model with 'age' without 'yoe'//
// this is better than model 6, bc with the smaller AIC //
reg os female avgself age age_sqr uae india asia europe bow_avgn bow_avgn2 ats_weighted qs_rankN qs_
```



```
//Report and Presentation Graphics
*graph twoway (scatter os avgself, symbol(d)) (lowess os avgself, bwidth(.99) lpattern(solid)), title("Overall Score vs Self-Rating")

*graph twoway (scatter os avgself, symbol(d)) (lfitci os avgself), title("Overall Score vs Self-Rating")


coefplot, drop(_cons) xline(0) xlabel(-75(25)75) xscale(range(-75(25)75)) title("Coefficients on Regessors in OLS Model")


*pnorm ehat, title("Normality of Residuals of Y=Overall Score on X=Self-Rating")

*graph twoway (scatter os avgself, symbol(d)), title("Overall Score vs Self-Rating")
/*
asdoc reg os female avgself age age_sqr uae india asia europe bow_avgn bow_avgn2 ats_weighted qs_rank qs_rank2 educsystem ats_uae ats_asi
estat ic
```

Classification Modelling Code



```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import roc_auc_score
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import roc_curve
from sklearn.metrics import r2_score
from sklearn.metrics import accuracy_score
```

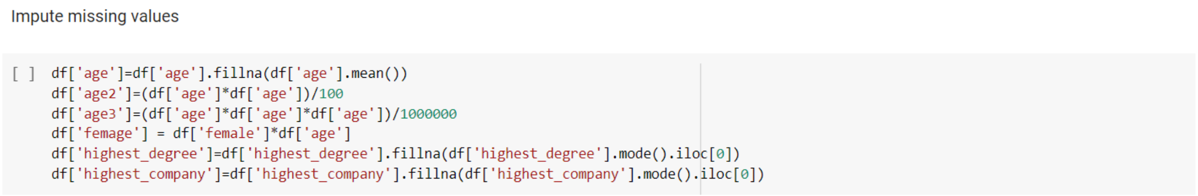
```
df = pd.read_csv('finaldata1 - Copy.csv')
```

Create interaction and polynomial terms

```
df['ats_weighted2'] = df['ats_weighted']**2
df['edu_ats'] = df['educsystem']*df['ats_weighted']
```

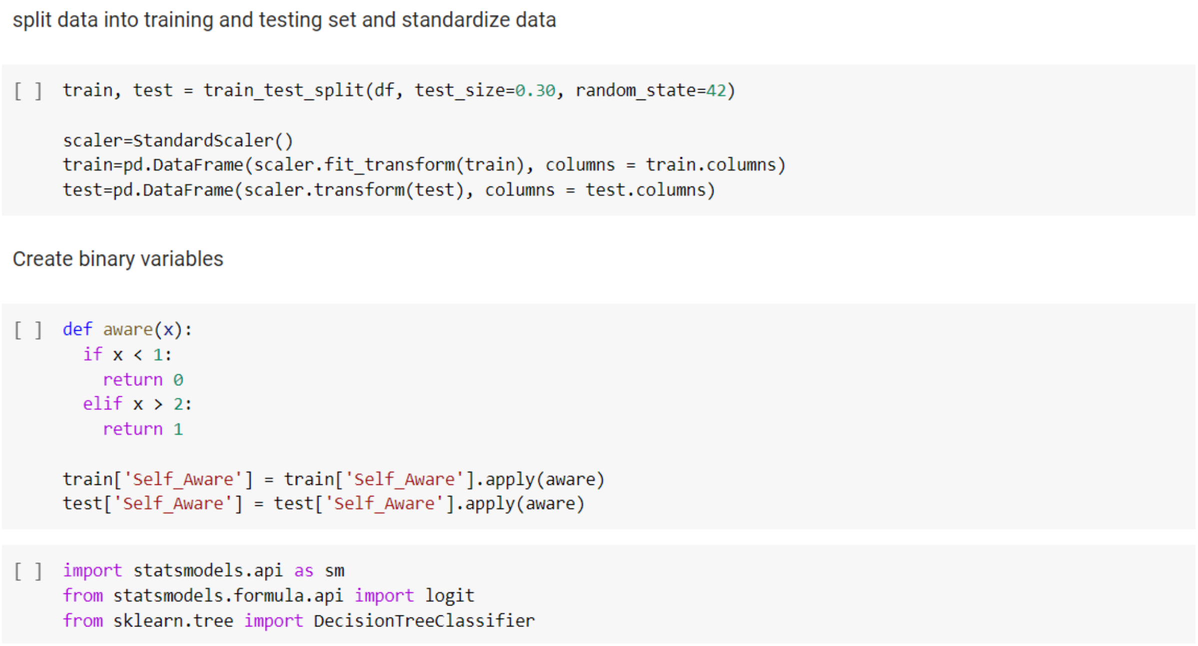
Define range for self-aware candidates to fall into

```
def confidence_label(x):
  if -3 <= x <= 3:
    return 1
  else:
    return 0

df['Self_Aware'] = df['selfawarenessfinal'].apply(confidence_label)
```



Impute missing values

```
df['age']=df['age'].fillna(df['age'].mean())
df['age2']=(df['age']*df['age'])/100
df['age3']=(df['age']*df['age']*df['age'])/1000000
df['femage'] = df['female']*df['age']
df['highest_degree']=df['highest_degree'].fillna(df['highest_degree'].mode().iloc[0])
df['highest_company']=df['highest_company'].fillna(df['highest_company'].mode().iloc[0])
```



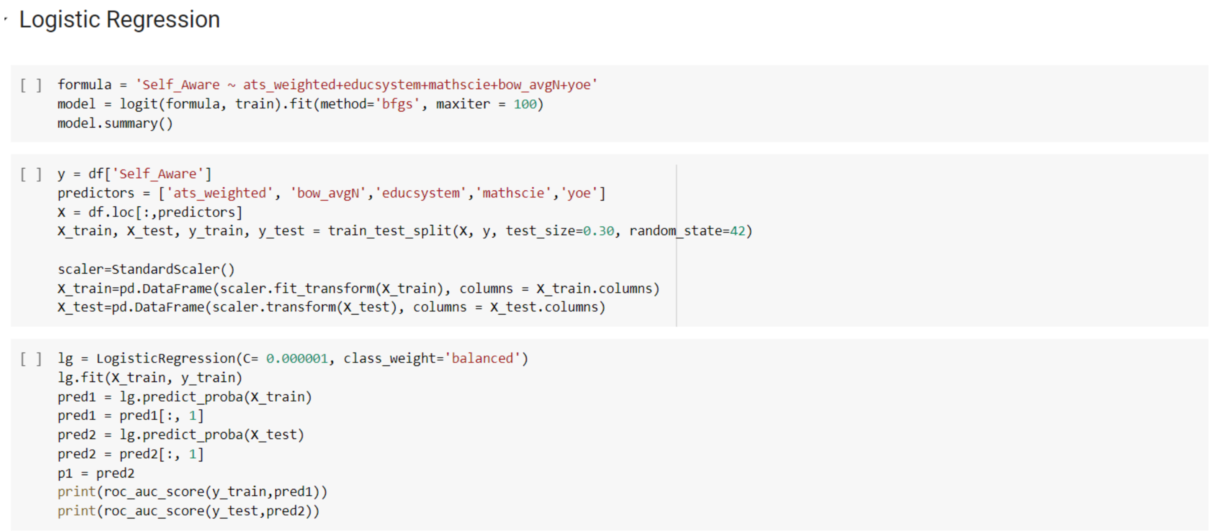
split data into training and testing set and standardize data

```
train, test = train_test_split(df, test_size=0.30, random_state=42)

scaler=StandardScaler()
train=pd.DataFrame(scaler.fit_transform(train), columns = train.columns)
test=pd.DataFrame(scaler.transform(test), columns = test.columns)
```

Create binary variables

```
def aware(x):
  if x < 1:
    return 0
  elif x > 2:
    return 1

train['Self_Aware'] = train['Self_Aware'].apply(aware)
test['Self_Aware'] = test['Self_Aware'].apply(aware)
```

```
import statsmodels.api as sm
from statsmodels.formula.api import logit
from sklearn.tree import DecisionTreeClassifier
```

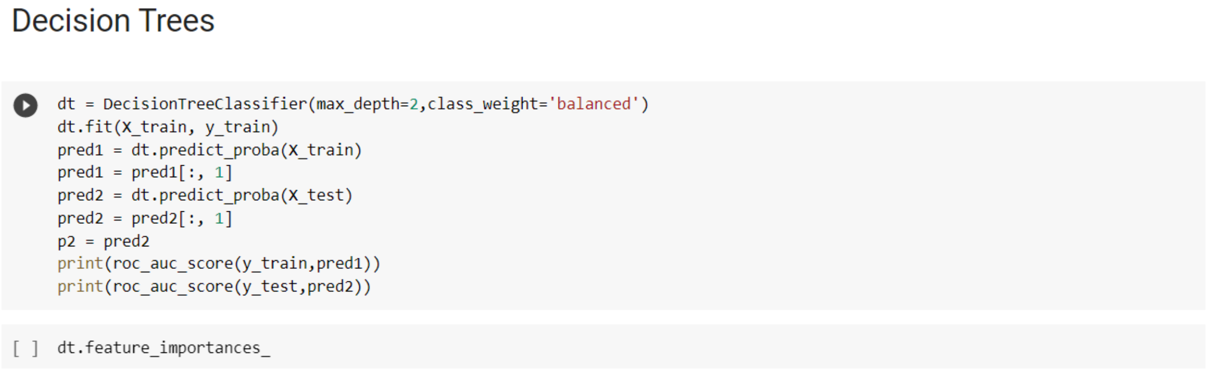


Logistic Regression

```
formula = 'Self_Aware ~ ats_weighted+educsystem+mathscie+bow_avgN+yoe'
model = logit(formula, train).fit(method='bfgs', maxiter = 100)
model.summary()
```

```
y = df['Self_Aware']
predictors = ['ats_weighted', 'bow_avgN','educsystem','mathscie','yoe']
X = df.loc[:,predictors]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30, random_state=42)

scaler=StandardScaler()
X_train=pd.DataFrame(scaler.fit_transform(X_train), columns = X_train.columns)
X_test=pd.DataFrame(scaler.transform(X_test), columns = X_test.columns)
```

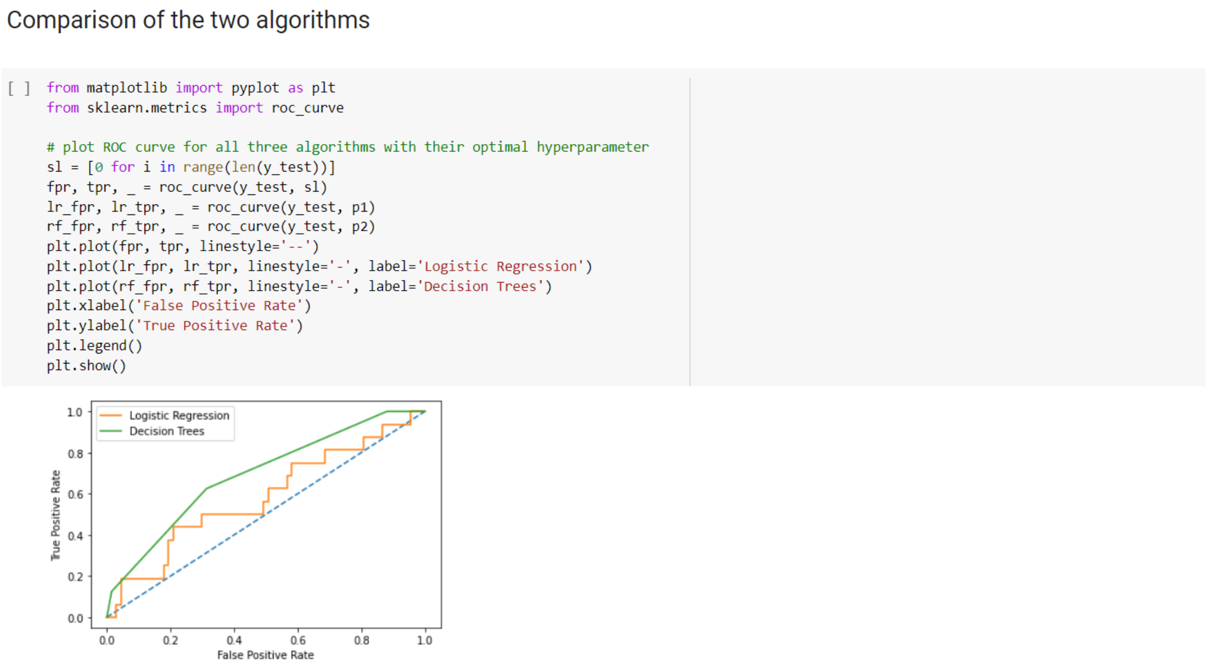
```
lg = LogisticRegression(C= 0.000001, class_weight='balanced')
lg.fit(X_train, y_train)
pred1 = lg.predict_proba(X_train)
pred1 = pred1[:, 1]
pred2 = lg.predict_proba(X_test)
pred2 = pred2[:, 1]
p1 = pred2
print(roc_auc_score(y_train,pred1))
print(roc_auc_score(y_test,pred2))
```



Decision Trees

```
dt = DecisionTreeClassifier(max_depth=2,class_weight='balanced')
dt.fit(X_train, y_train)
pred1 = dt.predict_proba(X_train)
pred1 = pred1[:, 1]
pred2 = dt.predict_proba(X_test)
pred2 = pred2[:, 1]
p2 = pred2
print(roc_auc_score(y_train,pred1))
print(roc_auc_score(y_test,pred2))
```

```
dt.feature_importances_
```

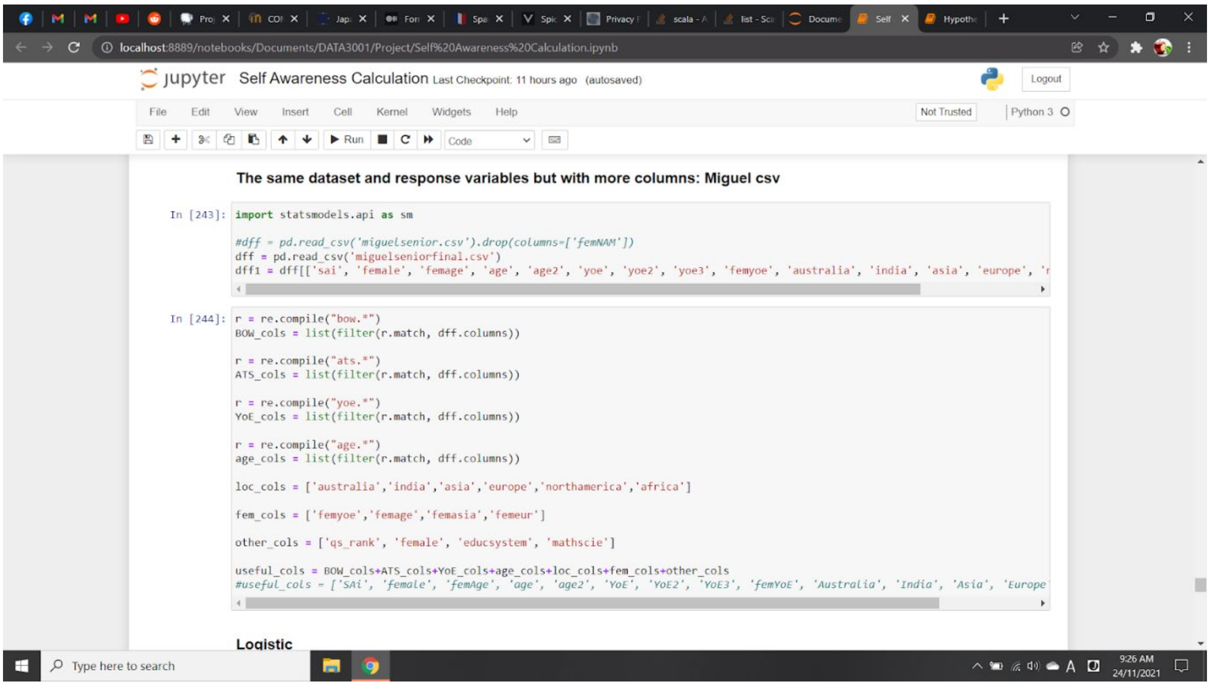


Comparison of the two algorithms

```
from matplotlib import pyplot as plt
from sklearn.metrics import roc_curve

# plot ROC curve for all three algorithms with their optimal hyperparameter
sl = [0 for i in range(len(y_test))]
fpr, tpr, _ = roc_curve(y_test, sl)
lr_fpr, lr_tpr, _ = roc_curve(y_test, p1)
rf_fpr, rf_tpr, _ = roc_curve(y_test, p2)
plt.plot(fpr, tpr, linestyle='--')
plt.plot(lr_fpr, lr_tpr, linestyle='-', label='Logistic Regression')
plt.plot(rf_fpr, rf_tpr, linestyle='-', label='Decision Trees')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.legend()
plt.show()
```

Method 2a Code

1. Preprocessing 2a



localhost:8889/notebooks/Documents/DATA3001/Project/Self%20Awareness%20Calculation.ipynb
jupyter Self Awareness Calculation Last Checkpoint: 11 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3
Run
Code
Type here to search
9:26 AM
24/11/2021

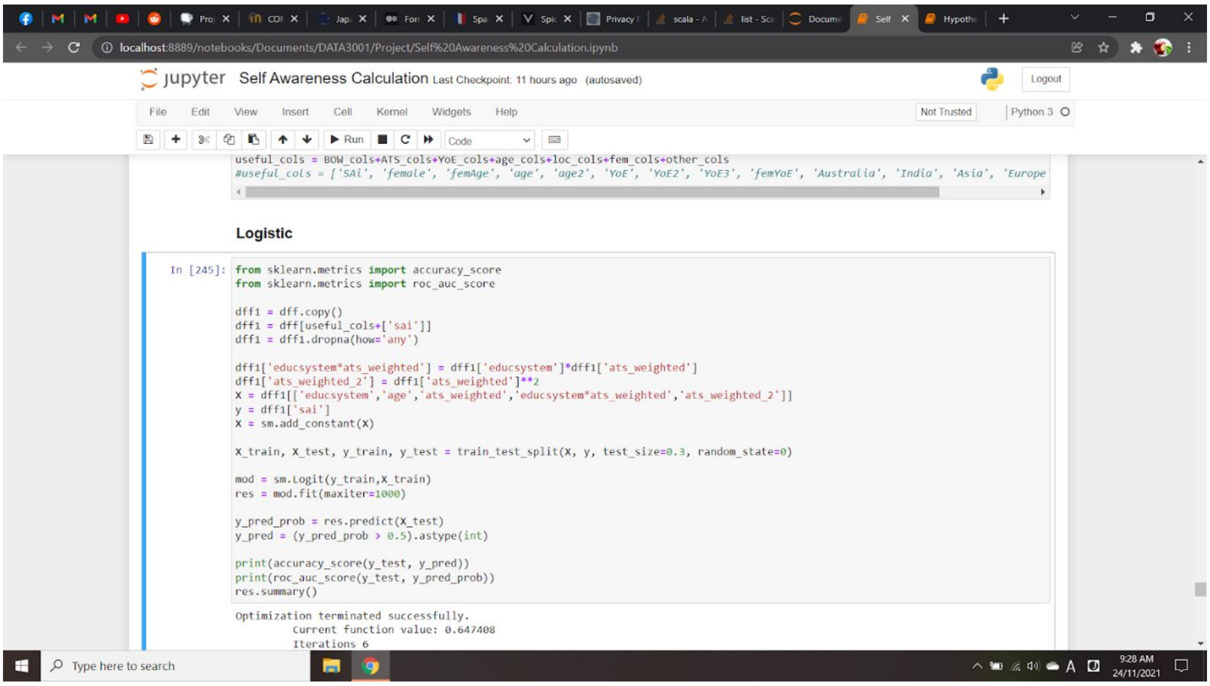
The same dataset and response variables but with more columns: Miguel csv

```
In [243]: import statsmodels.api as sm

          #dff = pd.read_csv('miguelsenior.csv').drop(columns=['femNAM'])
          dff = pd.read_csv('miguelseniorfinal.csv')
          dff1 = dff[['sai', 'female', 'femage', 'age', 'age2', 'yoe', 'yoe2', 'yoe3', 'femyoe', 'australia', 'india', 'asia', 'europe', 'r
```

```
In [244]: r = re.compile("bow.*")
          BOW_cols = list(filter(r.match, dff.columns))

          r = re.compile("ats.*")
          ATS_cols = list(filter(r.match, dff.columns))

          r = re.compile("yoe.*")
          YoE_cols = list(filter(r.match, dff.columns))

          r = re.compile("age.*")
          age_cols = list(filter(r.match, dff.columns))

          loc_cols = ['australia','india','asia','europe','northamerica','africa']

          fem_cols = ['femyoe','femage','femasia','femeur']

          other_cols = ['qs_rank', 'female', 'educsystem', 'mathscie']

          useful_cols = BOW_cols+ATS_cols+YoE_cols+age_cols+loc_cols+fem_cols+other_cols
          #useful_cols = ['SAi', 'female', 'femAge', 'age', 'age2', 'YoE', 'YoE2', 'YoE3', 'femYoE', 'Australia', 'India', 'Asia', 'Europe
```

Logistic

2. Logistic Regression 2a



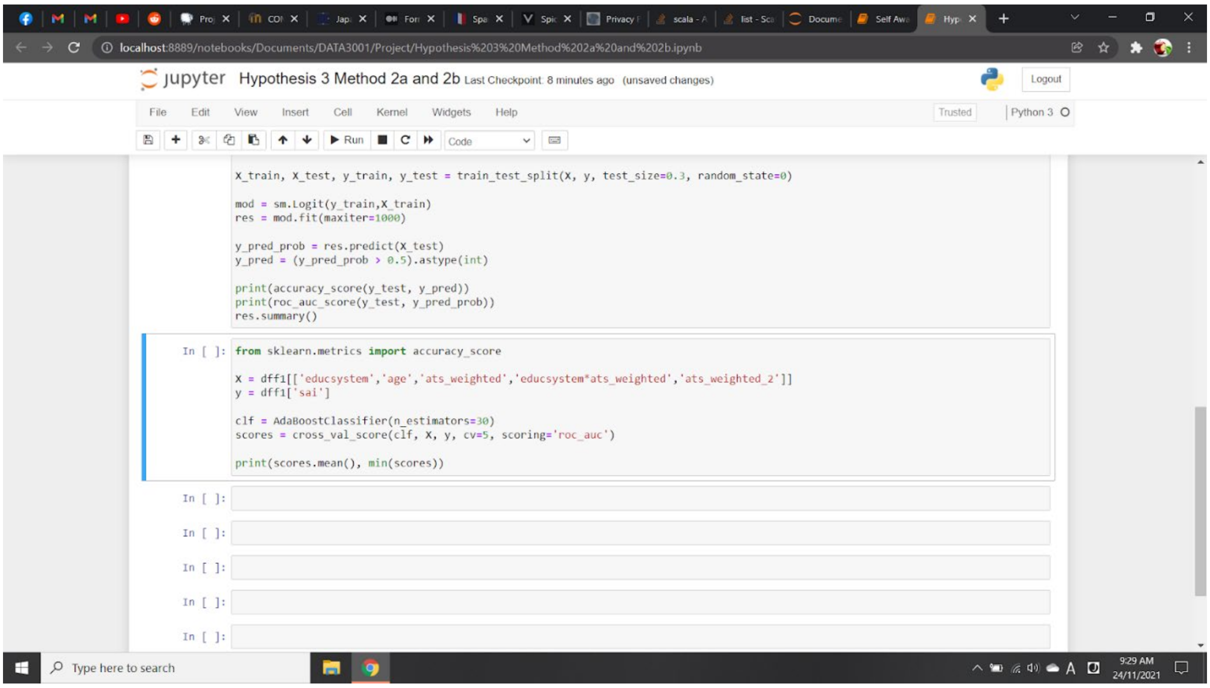
localhost:8889/notebooks/Documents/DATA3001/Project/Self%20Awareness%20Calculation.ipynb
jupyter Self Awareness Calculation Last Checkpoint: 11 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3
Run
Code
Type here to search
9:28 AM
24/11/2021

```
          useful_cols = BOW_cols+ATS_cols+YoE_cols+age_cols+loc_cols+fem_cols+other_cols
          #useful_cols = ['SAi', 'female', 'femAge', 'age', 'age2', 'YoE', 'YoE2', 'YoE3', 'femYoE', 'Australia', 'India', 'Asia', 'Europe
```

Logistic

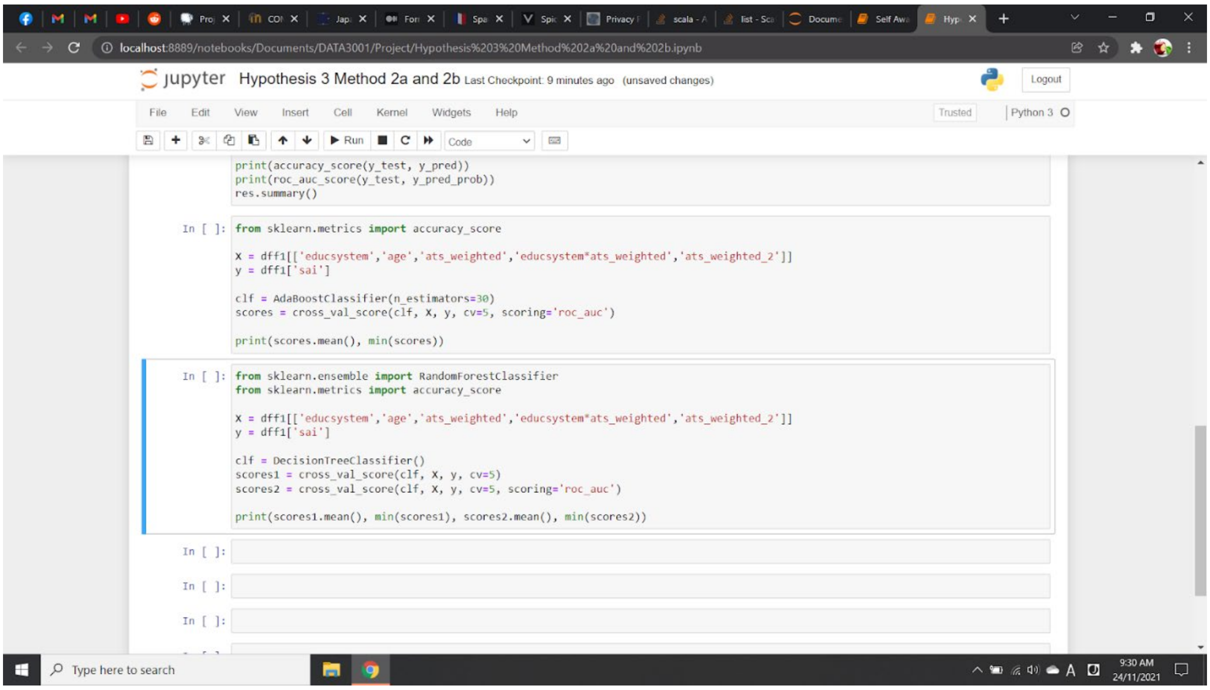
```
In [245]: from sklearn.metrics import accuracy_score
          from sklearn.metrics import roc_auc_score

          dff1 = dff.copy()
          dff1 = dff[useful_cols+['sai']]
          dff1 = dff1.dropna(how='any')

          dff1['educsystem*ats_weighted'] = dff1['educsystem']*dff1['ats_weighted']
          dff1['ats_weighted_2'] = dff1['ats_weighted']**2
          X = dff1[['educsystem','age','ats_weighted','educsystem*ats_weighted','ats_weighted_2']]
          y = dff1['sai']
          X = sm.add_constant(X)

          X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)

          mod = sm.Logit(y_train,X_train)
          res = mod.fit(maxiter=1000)

          y_pred_prob = res.predict(X_test)
          y_pred = (y_pred_prob > 0.5).astype(int)

          print(accuracy_score(y_test, y_pred))
          print(roc_auc_score(y_test, y_pred_prob))
          res.summary()

          Optimization terminated successfully.
                   Current function value: 0.647408
                   Iterations 6
```

3. Adaboost 2a


localhost:8889/notebooks/Documents/DATA3001/Project/Hypothesis%203%20Method%202a%20and%202b.ipynb
jupyter Hypothesis 3 Method 2a and 2b Last Checkpoint: 8 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
Type here to search
9:29 AM
24/11/2021

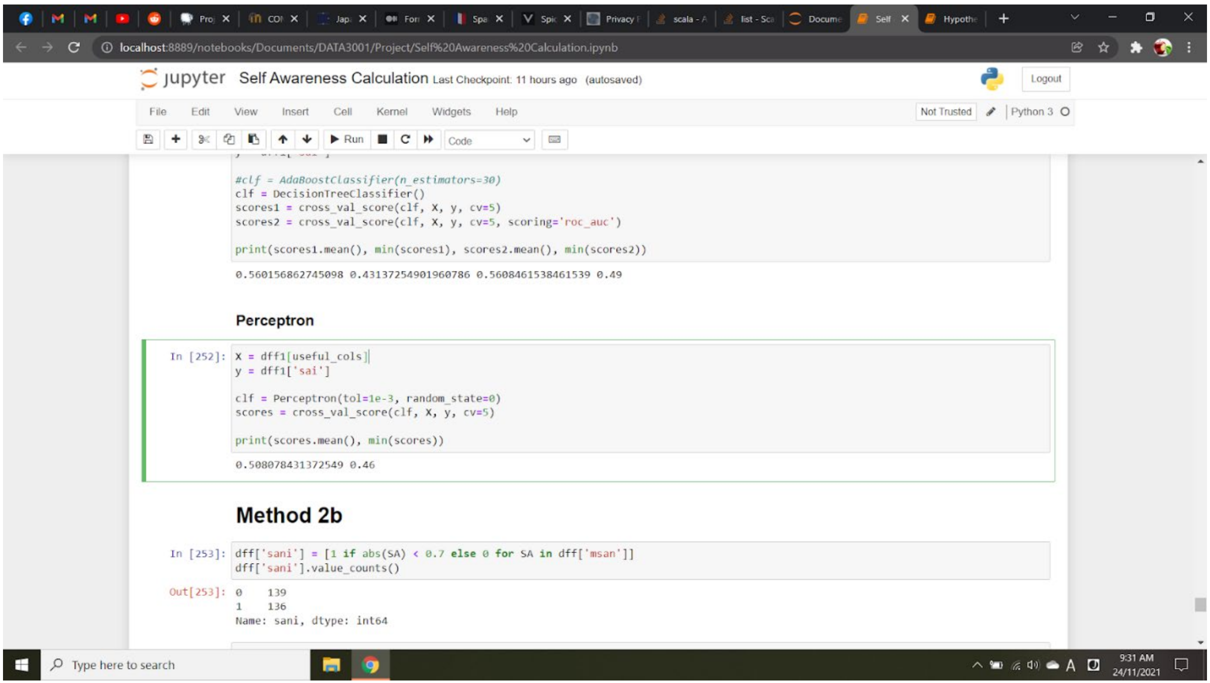
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)

mod = sm.Logit(y_train,X_train)
res = mod.fit(maxiter=1000)

y_pred_prob = res.predict(X_test)
y_pred = (y_pred_prob > 0.5).astype(int)

print(accuracy_score(y_test, y_pred))
print(roc_auc_score(y_test, y_pred_prob))
res.summary()
```

```
In [ ]: from sklearn.metrics import accuracy_score

        X = dff1[['educsystem','age','ats_weighted','educsystem*ats_weighted','ats_weighted_2']]
        y = dff1['sai']

        clf = AdaBoostClassifier(n_estimators=30)
        scores = cross_val_score(clf, X, y, cv=5, scoring='roc_auc')

        print(scores.mean(), min(scores))

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:
```


localhost:8889/notebooks/Documents/DATA3001/Project/Hypothesis%203%20Method%202a%20and%202b.ipynb
jupyter Hypothesis 3 Method 2a and 2b Last Checkpoint: 9 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
Type here to search
9:30 AM
24/11/2021

```
print(accuracy_score(y_test, y_pred))
print(roc_auc_score(y_test, y_pred_prob))
res.summary()

In [ ]: from sklearn.metrics import accuracy_score

        X = dff1[['educsystem','age','ats_weighted','educsystem*ats_weighted','ats_weighted_2']]
        y = dff1['sai']

        clf = AdaBoostClassifier(n_estimators=30)
        scores = cross_val_score(clf, X, y, cv=5, scoring='roc_auc')

        print(scores.mean(), min(scores))

In [ ]: from sklearn.ensemble import RandomForestClassifier
        from sklearn.metrics import accuracy_score

        X = dff1[['educsystem','age','ats_weighted','educsystem*ats_weighted','ats_weighted_2']]
        y = dff1['sai']

        clf = DecisionTreeClassifier()
        scores1 = cross_val_score(clf, X, y, cv=5)
        scores2 = cross_val_score(clf, X, y, cv=5, scoring='roc_auc')

        print(scores1.mean(), min(scores1), scores2.mean(), min(scores2))

In [ ]:

In [ ]:

In [ ]:
```

4. Perceptron 2a


localhost:8889/notebooks/Documents/DATA3001/Project/Self%20Awareness%20Calculation.ipynb
jupyter
Self Awareness Calculation Last Checkpoint: 11 hours ago (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3
Run
Code
Type here to search
9:31 AM
24/11/2021

```
#clf = AdaBoostClassifier(n_estimators=30)
clf = DecisionTreeClassifier()
scores1 = cross_val_score(clf, X, y, cv=5)
scores2 = cross_val_score(clf, X, y, cv=5, scoring='roc_auc')

print(scores1.mean(), min(scores1), scores2.mean(), min(scores2))
```

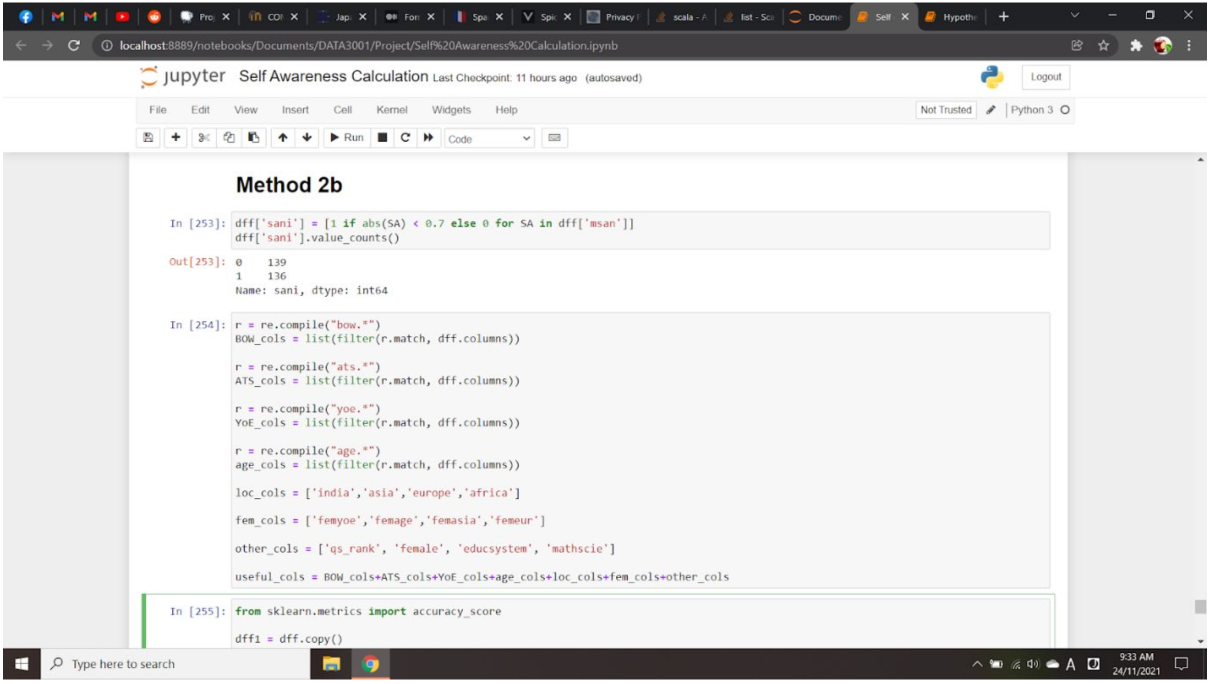
```
0.560156862745098 0.43137254901960786 0.5608461538461539 0.49
```

Perceptron

```
In [252]: X = dff1[useful_cols]
          y = dff1['sai']

          clf = Perceptron(tol=1e-3, random_state=0)
          scores = cross_val_score(clf, X, y, cv=5)

          print(scores.mean(), min(scores))

          0.508078431372549 0.46
```

Method 2b

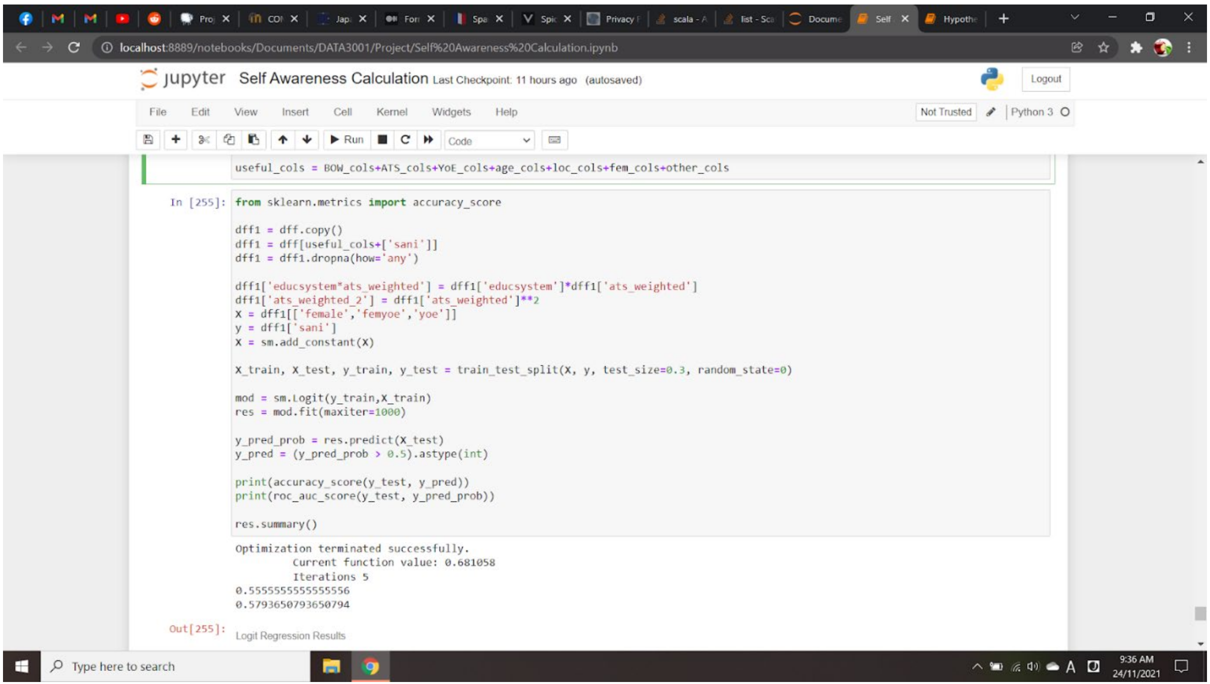
```
In [253]: dff['sani'] = [1 if abs(SA) < 0.7 else 0 for SA in dff['msan']]
          dff['sani'].value_counts()

Out[253]: 0    139
          1    136
          Name: sani, dtype: int64
```


localhost:8889/notebooks/Documents/DATA3001/Project/Self%20Awareness%20Calculation.ipynb
jupyter
Self Awareness Calculation Last Checkpoint: 11 hours ago (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3
Run
Code
Type here to search
9:33 AM
24/11/2021

Method 2b

```
In [253]: dff['sani'] = [1 if abs(SA) < 0.7 else 0 for SA in dff['msan']]
          dff['sani'].value_counts()

Out[253]: 0    139
          1    136
          Name: sani, dtype: int64
```

```
In [254]: r = re.compile("bow.*")
          BOW_cols = list(filter(r.match, dff.columns))

          r = re.compile("ats.*")
          ATS_cols = list(filter(r.match, dff.columns))

          r = re.compile("yoe.*")
          YoE_cols = list(filter(r.match, dff.columns))

          r = re.compile("age.*")
          age_cols = list(filter(r.match, dff.columns))

          loc_cols = ['india','asia','europe','africa']

          fem_cols = ['femyoe','femage','femasia','femeur']

          other_cols = ['qs_rank', 'female', 'educsystem', 'mathscie']

          useful_cols = BOW_cols+ATS_cols+YoE_cols+age_cols+loc_cols+fem_cols+other_cols

In [255]: from sklearn.metrics import accuracy_score

          dff1 = dff.copy()
```


localhost:8889/notebooks/Documents/DATA3001/Project/Self%20Awareness%20Calculation.ipynb
jupyter
Self Awareness Calculation Last Checkpoint: 11 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3
Run
Code
Type here to search
9:36 AM
24/11/2021

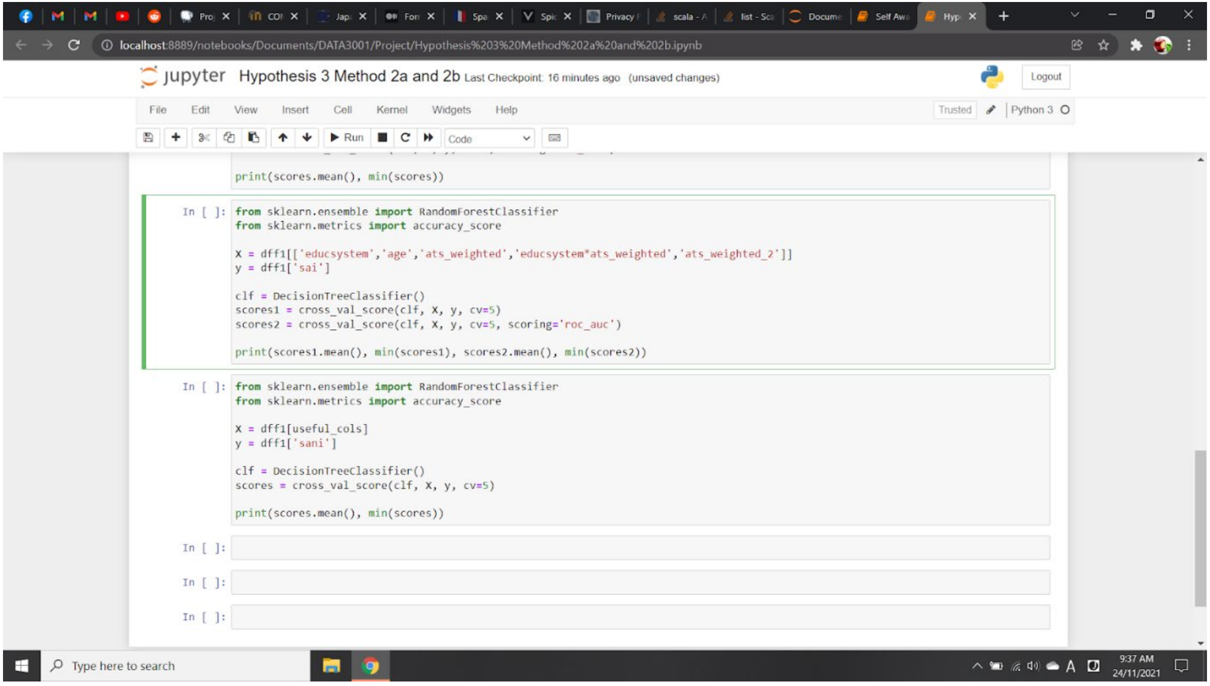
```
useful_cols = BOW_cols+ATS_cols+YoE_cols+age_cols+loc_cols+fem_cols+other_cols
```

In [255]:

```
from sklearn.metrics import accuracy_score

dff1 = dff.copy()
dff1 = dff[useful_cols+['sani']]
dff1 = dff1.dropna(how='any')

dff1['educsystem*ats_weighted'] = dff1['educsystem']*dff1['ats_weighted']
dff1['ats_weighted_2'] = dff1['ats_weighted']**2
X = dff1[['female','femyoe','yoe']]
y = dff1['sani']
X = sm.add_constant(X)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)

mod = sm.Logit(y_train,X_train)
res = mod.fit(maxiter=1000)

y_pred_prob = res.predict(X_test)
y_pred = (y_pred_prob > 0.5).astype(int)

print(accuracy_score(y_test, y_pred))
print(roc_auc_score(y_test, y_pred_prob))

res.summary()
```

```
Optimization terminated successfully.
         Current function value: 0.681058
         Iterations 5
0.5555555555555556
0.5793650793650794
```

Out[255]: Logit Regression Results


localhost:8889/notebooks/Documents/DATA3001/Project/Hypothesis%203%20Method%202a%20and%202b.ipynb
jupyter
Hypothesis 3 Method 2a and 2b Last Checkpoint: 16 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
Type here to search
9:37 AM
24/11/2021

```
print(scores.mean(), min(scores))
```

In [ ]:

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score

X = dff1[['educsystem','age','ats_weighted','educsystem*ats_weighted','ats_weighted_2']]
y = dff1['sai']

clf = DecisionTreeClassifier()
scores1 = cross_val_score(clf, X, y, cv=5)
scores2 = cross_val_score(clf, X, y, cv=5, scoring='roc_auc')

print(scores1.mean(), min(scores1), scores2.mean(), min(scores2))
```

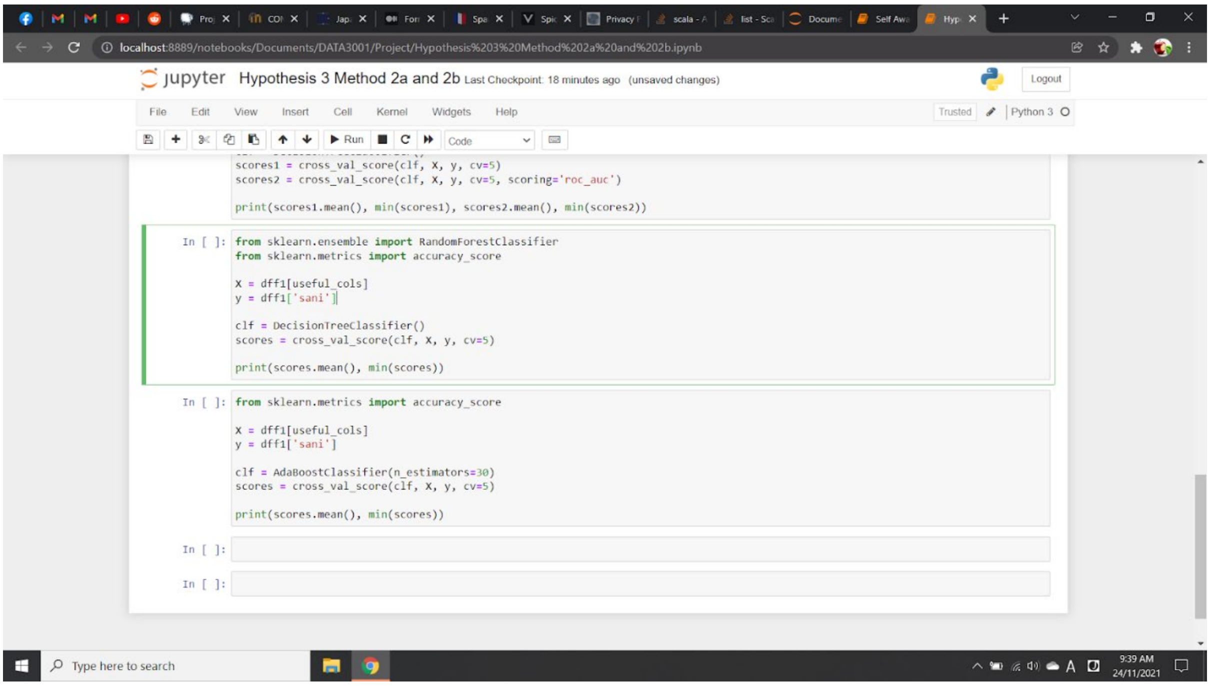
In [ ]:

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score

X = dff1[useful_cols]
y = dff1['sani']

clf = DecisionTreeClassifier()
scores = cross_val_score(clf, X, y, cv=5)

print(scores.mean(), min(scores))
```

In [ ]:

In [ ]:

In [ ]:



localhost:8889/notebooks/Documents/DATA3001/Project/Hypothesis%203%20Method%202a%20and%202b.ipynb
jupyter Hypothesis 3 Method 2a and 2b Last Checkpoint: 18 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
Type here to search
9:39 AM
24/11/2021

```
scores1 = cross_val_score(clf, X, y, cv=5)
scores2 = cross_val_score(clf, X, y, cv=5, scoring='roc_auc')

print(scores1.mean(), min(scores1), scores2.mean(), min(scores2))
```

In [ ]:

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score

X = dff1[useful_cols]
y = dff1['sani']

clf = DecisionTreeClassifier()
scores = cross_val_score(clf, X, y, cv=5)

print(scores.mean(), min(scores))
```

In [ ]:

```
from sklearn.metrics import accuracy_score

X = dff1[useful_cols]
y = dff1['sani']

clf = AdaBoostClassifier(n_estimators=30)
scores = cross_val_score(clf, X, y, cv=5)

print(scores.mean(), min(scores))
```

In [ ]:

In [ ]: