UNIVERSITY OF KWAZULU-NATAL ™

INYUVESI YAKWAZULU-NATALI

STAT305: Biostatistics Methods

Assignment 2

Topic: Data Scientist Salary

SAS Report and Analysis

Group A

219095164	Akita Sheodin
214517334	Yasthika Ramgobin
217005481	Kaylashni Moodley
218002141	Minenhle Beloved Sethebe
219004186	Noluvo Nobuhle Mathaba

We, the group members of Group A, hereby declare that Assignment 2 was completed by all group members over multiple zoom meetings.

Table of Contents

1. Aim

With the rise of the 4th industrial revolution, data science has become an essential part of many industries in the modern world, hence making data scientists high in demand globally. There are several factors that influence a data scientist's salary such as the software skills required, education level, experience, and type of industry, to name a few. Accordingly, through the implementation of association tests and logistic regression, the aim of this study is to determine which factors of the given dataset have a significant association with the salary of a data scientist.

2. Introduction to dataset

The dataset used for this study is data scientist salary data, which comprises of 212 observations (rows) and 12 variables. Each row in the dataset pertains to an individual data scientist and their characteristics (e.g., age, degree, salary etc.). The variables in the dataset are a mixture of categorical (qualitative) and numerical (quantitative) variables.

The main variable of interest (response) for this study is Salary, which is a categorical variable with binary outcomes, that describes whether the individual's salary is below average or above average. Age is a quantitative variable describing the individual's age in years. Company Rating is a numerical variable that describes the rating of the company the data scientist works for. Founded is also a numeric variable describing the year in which the company was founded.

Degree and Type of Ownership are classified as categorical variables. Furthermore, both Company Size (the range of the number of employees working in the company) and Revenue (the total revenue of the company per year) are categorical variables, since they comprise of ranges of values in the data. Lastly, the variables Python, SAS, Excel and SQL specify whether that software skill is required for the position and each of them represent a categorical variable with binary outcomes (Yes/No) in the dataset.

3. Variables

3.1. Measurement scale

There are four basic levels of measurement, namely: nominal, ordinal, interval, and ratio. The measurement scale for each variable in the dataset have been identified as follows:
The variables Type of Ownership, Python, SAS, Excel and SQL have a nominal scale of measurement since they are categorical in nature and do not have ordered categories. The variables Salary, Degree and Company Size have an ordinal scale of measurement since they both have ordered categories or categories that can be ranked.

For the variable Revenue, however, one of the categories are labelled as 'Unknown/Non-Applicable'. A two-way table for Company Size by Revenue, showed that most of the data scientists in the sample who had Revenue marked as 'Unknown/Non-Applicable' worked for companies with a size of 1000 employees or less. Hence, the assumption was that smaller companies will generally generate less revenue than larger companies, thus 'Unknown/Non-Applicable' was ranked as the lowest category, followed by the $5-$100 million category, then

$100 million-$2 billion and lastly $2-$10+ billion, thus making the variable Revenue ordinal. The variable Founded has an interval scale of measurement since it does not have a true/meaningful zero. Further, the variables Age and Company Rating have a ratio scale of measurement since they contain a true zero point.

3.2. One-way tables: Categorical variables

The one-way tables for the categorical variables in the dataset can be found in Appendix 1. From these outputs, we see that 44.34% of the data scientists in the sample are earning a salary that is above average and the remaining 55.66% earn a salary that is below average. Hence, there is a higher proportion of data scientists that earn a below average salary. In addition, almost half (49.06%) of data scientists in the sample have no degree qualification. Privately owned companies seem to have the highest proportion of data scientists (45.28%), compared to the other types of ownership in the dataset. In terms of the software skills required for their position, there seems to be a higher proportion of data scientists (54.72% and 53.30%) that require excel and sql, respectively. Whereas there is a higher proportion of data scientists (55.19% and 87.74%) that do not require python and sas, respectively, as a skill.

3.3. Summary measures: Quantitative variables

The output for the basic summary measures of the quantitative variables can be found in Appendix 2. The measures of central tendency were obtained to describe the key characteristics of the dataset and identify the central position (location) in the dataset.

The mean obtained for Age is 40.1557, implying the average age of a data scientist in this dataset is approximately 41 years old. Furthermore, the median is 39 and the mode is 25, indicating that data scientists aged 25 occur most frequently in the data. The age data is slightly positively skewed, as the mean is greater than the median. The standard deviation is 11.2397 and the variance is 126.3311, suggesting that age has a high variability. The oldest data scientist in the sample is 60 years, and the youngest is 25 years, hence the range is estimated to be 35 and the interquartile range is 21.5. The measure of spread indicates that the age data has a significant spread or variance from the center of the data.

The measures of central tendency obtained for Company Rating indicate that the average rating of a company is 3.6236, the median is 3.70 and the mode is 3.80. This suggests that company rating has central tendency around 3.0 and that the data is slightly negatively skewed, as the mean is less than the median. The standard deviation is 0.5518 and variance is 0.3045, indicating that company rating has a low variability and is more stable. The maximum rating is 4.7 and the minimum is 2.2, hence the range is 2.5 and the interquartile range is 0.6. This suggests that the company rating data is less spread/clustered together.

The mean obtained for Founded (in years) is 1980.844 ($\approx$ 1981), the median is 1982 and the mode is 1996. This indicates that, on average, data scientists are working for companies that were founded in the year 1981. The data is slightly negatively skewed as the mean is less than the median. The standard deviation is 11.2397 and variance is 126.3311, therefore the dataset has a relatively high variability. The maximum and minimum values for founded explain that

the earliest founded company was in 1961 and the most recent founded company was 1996 in the dataset. The range is estimated to be 35 and the interquartile range is 21.5, signifying that there is a great dispersion for founded and observations are spread out.

4. Tests of association

4.1. Relevant categorical variables of interest

As per the one-way tables produced above, the relevant categorical variables of interest in this study, based on those that did not have many categories/levels, were: Degree, Revenue, Python, SAS, Excel and SQL. For the variables Salary, Degree and Revenue, it was evident that their respective categories were ranked, however it was not presented as such in the initial one-way tables, hence they were ordered accordingly in SAS before performing the tests of association.

4.2. Two-way tests of association

The following tests of association were performed between each relevant categorical explanatory variable and the response variable of interest (Salary) at a 5% level of significance. The SAS output for each test can be found in Appendix 3.

a) Test of association between Degree and Salary:

Since degree and salary are both ordinal, the following *correlation test* was performed:

H_0: There is no linear association (zero correlation) between degree and salary.

H_1: There is a linear association (nonzero correlation) between degree and salary.

$$Q_{CS} = 15.2282 \sim \chi^2_{(1)}$$

$$p - value < 0.0001$$

Since the p-value < 0.05, H_0 is rejected at a 5% level of significance. Thus, there is sufficient evidence to conclude that there is a significant linear association between the degree type of the individual and their salary.

b) Test of association between Revenue and Salary:

Since revenue and salary are both ordinal, the following *correlation test* was performed:

H_0: There is no linear association (zero correlation) between revenue and salary.

H_1: There is a linear association (nonzero correlation) between revenue and salary.

$$Q_{CS} = 3.7741 \sim \chi^2_{(1)}$$

$$p - value = 0.0521$$

Since the p-value = 0.0521 > 0.05, the null hypothesis H_0 is not rejected at a 5% level of significance. Thus, there is insufficient evidence to conclude that there is a linear association between the total annual revenue of the company and the salary of a data scientist.

c) Test of association between Python and Salary:

Since python is nominal and salary is ordinal (where both variables are binary), the following *general association test* was performed:

H_0: There is no association between python and salary.

H_1: There is an association between python and salary.

$$Q = 36.8150 \sim \chi^2_{(1)}$$

$$p - value < 0.0001$$

Since the p-value < 0.05, H_0 is rejected at a 5% level of significance. Thus, there is sufficient evidence to conclude that there is a significant association between a data scientist requiring Python as a software skill for their job and the salary they earn.

d) Test of association between SAS and Salary:

Since sas is nominal and salary is ordinal (where both variables are binary), the following *general association test* was performed:

H_0: There is no association between sas and salary.

H_1: There is an association between sas and salary.

$$Q = 0.3829 \sim \chi^2_{(1)}$$

$$p - value = 0.5361$$

Since the p-value = 0.5361 > 0.05, the null hypothesis H_0 is not rejected at a 5% level of significance. Thus, there is insufficient evidence to conclude that there is an association between a data scientist requiring SAS as a software skill for their job and the salary they earn.

e) Test of association between Excel and Salary:

Since excel is nominal and salary is ordinal (where both variables are binary), the following *general association test* was performed:

H_0: There is no association between excel and salary.

H_1: There is an association between excel and salary.

$$Q = 0.9053 \sim \chi^2_{(1)}$$

$$p - value = 0.3414$$

Since the p-value = 0.3414 > 0.05, the null hypothesis H_0 is not rejected at a 5% level of significance. Thus, there is insufficient evidence to conclude that there is an association between a data scientist requiring Excel as a software skill for their job and the salary they earn.

f) Test of association between SQL and Salary:

Since sql is nominal and salary is ordinal (where both variables are binary), the following *general association test* was performed:

H_0: There is no association between sql and salary.

H_1: There is an association between sql and salary.

$$Q = 4.7651 \sim \chi^2_{(1)}$$

$$p - value = 0.0290$$

Since the p-value = 0.0290 < 0.05, H_0 is rejected at a 5% level of significance. Thus, there is sufficient evidence to conclude that there is a significant association between a data scientist requiring SQL as a software skill for their job and the salary they earn.

5. Fitting a statistical model

5.1. Logistic regression model

A logistic regression model for the data, which included all the relevant categorical explanatory variables and the quantitative explanatory variables, was fitted as follows:

$$\begin{aligned} logit(\hat{y}) = &-2.2782 + 1.9927x_1 + 1.2174x_2 - 2.1238x_3 - 2.0462x_4 - 0.7535x_5 \\ &+ 1.6157x_6 - 0.00308x_7 - 0.6681x_8 + 0.3583x_9 - 0.0217x_{10} \\ &+ 0.8170x_{11} \end{aligned}$$

Where:

$\hat{y}$ is the predicted salary (coded as 1 for above average and 0 for below average); x_1 is 1 if the individual has a postgraduate degree and 0 otherwise; x_2 is 1 if the individual has an undergraduate degree and 0 otherwise; x_3 is 1 if the company's revenue is between \$100 million to \$2 billion and 0 otherwise; x_4 is 1 if the company's revenue is between \$5 to \$100 million and 0 otherwise; x_5 is 1 if the company's revenue is unknown/non-applicable and 0 otherwise; x_6 is 1 if python is required for the position and 0 otherwise; x_7 is 1 if sas is required for the position and 0 otherwise; x_8 is 1 if excel is required for the position and 0 otherwise; x_9 is 1 if sql is required for the position and 0 otherwise; x_{10} is the age of the individual in years and x_{11} is the rating of the company the individual works for.

5.2. Comparing the significance of the categorical explanatory variables

The tests of association concluded that the variables Degree, Python and SQL have a significant association with the response variable of interest (Salary). Whereas the logistic regression model showed that Degree, Revenue and Python were the significant variables when predicting the response variable. The reason for differences in the results is because when performing tests of association, we use a contingency table that examines the effect of one explanatory variable on the response. However, when using a generalized linear model, such as the logistic regression model, it is possible to simultaneously analyse the effect of multiple explanatory variables on the response variable. Moreover, the tests of association are performed by using categorical variables only, whereas the logistic regression model can incorporate both continuous and categorical variables, hence leading to differences in the results.

6. The best fitting model

6.1. Reduced logistic regression model

A reduced logistic regression model was fitted by incorporating only the significant predictor variables (i.e., variables with a p-value < 0.05) into the model:

$$logit(\hat{y}) = -3.5724 + 1.8884x_1 + 1.1784x_2 - 2.2926x_3 - 2.1127x_4 - 0.9107x_5 + 1.6910x_6 + 0.9128x_7$$

Where:

$\hat{y}$ is the predicted salary (coded as 1 for above average and 0 for below average); x_1 is 1 if the individual has a postgraduate degree and 0 otherwise; x_2 is 1 if the individual has an undergraduate degree and 0 otherwise; x_3 is 1 if the company's revenue is between \$100 million to \$2 billion and 0 otherwise; x_4 is 1 if the company's revenue is between \$5 to \$100 million and 0 otherwise; x_5 is 1 if the company's revenue is unknown/non-applicable and 0 otherwise; x_6 is 1 if python is required for the position and 0 otherwise and x_7 is the rating of the company the individual works for.

The first model fitted was a logistic regression model with all the relevant categorical and quantitative explanatory variables in the data (i.e., Degree, Revenue, Python, SAS, Excel, SQL, Age, Company Rating and Founded). Thereafter, a reduced logistic regression model was fitted which retained only the significant predictor variables from the initial logistic regression model (i.e., Degree, Revenue, Python and Company Rating). To compare the two models and find the best fitted model, the Akaike Information Criterion (AIC) was employed as a criterion to assess the model fit and this was an appropriate measure to use as it accounts for the varying number of predictors in the model. For both models, we found that the AIC value was lower for the intercept and covariates model in comparison to the intercept only model, hence both fitted models provided a better fit to the data than their respective intercept only models.

Moreover, a test of the global null hypothesis, which tests the overall fit of the model, ($H_0: \beta_1 = \beta_2 \ldots = \beta_k = 0 \; vs. \; H_1: At \; least \; one \; \beta_j \neq 0$) was rejected for both models, as all 3 tests gave significant p-values, and so we are able to deduce that at least one of the predictor variables in the model has a significant effect on the response.

The first model had an AIC value of 226.462, while the second model had an AIC value of 223.923, hence the *reduced logistic regression model* was found to be the best fitting model as it produced the smallest AIC. In addition, the reduced logistic regression model had a concordance index of 0.837 which is quite high, indicating that the model has a good ability to predict the outcomes of the response variable, Salary.

6.2. Hosmer and Lemeshow goodness-of-fit test for the final model

H_0: The fitted model is adequate

H_1: The fitted model is not adequate

$$\chi^2 = 7.9455 \sim \chi^2_{(8)}$$

$$p - value = 0.4388$$

Since the p-value = 0.4388 > 0.05, the null hypothesis H_0 is not rejected at a 5% level of significance. Hence, it can be concluded that the fitted model is adequate for the data i.e., the model is fitting the data well.

6.3. Overdispersion check

The final model's deviance divided by its degrees of freedom yields a value of 1.7724, as seen in Appendix 5. This value is greater than 1, indicating that there is a bit of overdispersion present in the data. In order to correct for the overdispersion in the data, the dispersion parameter (Φ) is estimated by the Pearson Chi-square statistic divided by the degrees of freedom (Φ = 1.3015), hence the additional variability in the data is accounted for.

7. Results and interpretation

7.1. Odds ratios and confidence intervals

The event of interest for the logistic regression was an above average salary. The odds ratio estimates ($\widehat{OR}$) for the final model can be found in Appendix 6, where each OR estimate was given by $e^{\hat{\beta}_j}$.

For the categorical predictor Degree, all degree groups have a significantly higher likelihood of an above average salary compared to the no degree group. For postgraduate degree, the $\widehat{OR}$ = 6.609 > 1, with the reference category being 'None' i.e. no degree. This indicates that data scientists with a postgraduate degree have a higher likelihood of earning an above average salary compared to those with no degree. The 95% confidence interval (CI) for the OR is given by (2.111; 20.689), and since 1 is not contained in the interval, we can conclude that there is a significant difference in the odds of an above average salary between those with a postgraduate degree and those without a degree. Furthermore, the entire interval is greater than 1, indicating that postgraduate data scientists have a significantly higher likelihood of an above average salary than those data scientists with no degree.

The $\widehat{OR}$ = 3.249 indicates that data scientists with an undergraduate degree have a higher likelihood of earning an above average salary compared to those with no degree. This is confirmed by the 95% CI of (1.481; 7.127) > 1, which also indicates that there is a significant difference in the odds of an above average salary between data scientists with an undergraduate degree and those without a degree.

For the categorical predictor Revenue, all the revenue groups have a lower likelihood of an above average salary in comparison to the reference category, where revenue is between $2 to $10+ billion (i.e. the highest revenue group). More specifically, the $\widehat{OR}$ = 0.101 < 1 for revenue between $100 to $2 billion, indicating that data scientists working in companies with a revenue of $100 to $2 billion have a lower likelihood of an above average salary than those working in companies with a higher level of revenue. This result is confirmed by the 95% CI of (0.037; 0.273), which is less than 1. Moreover, 1 is not contained in the interval, hence there is a significant difference in the odds of an above average salary between these two revenue groups.

Similarly, for revenue between \$5 to \$100 million, the $\widehat{OR}$ = 0.121 < 1, indicating that data scientists working in companies with a revenue of \$5 to \$100 million have a lower likelihood of an above average salary than data scientists whose company revenue is between \$2 to \$10+ billion. Accordingly, this is supported by the 95% CI of (0.039; 0.372) < 1, and since 1 is not in the interval, there is a significant difference in the odds of an above average salary between these two revenue groups.

Further, the $\widehat{OR}$ = 0.402 indicates that data scientists with an unknown/non-applicable revenue have a lower likelihood of an above average salary than data scientists whose company revenue is between \$2 to \$10+ billion. The 95% CI of (0.135; 1.202) contains 1, implying there is no significant difference in the odds of an above average salary between these two revenue groups.

Additionally, for the categorical predictor Python, the $\widehat{OR}$ = 5.425 > 1, which implies that the data scientists that require the python software skill have a higher likelihood of earning an above average salary than those who do not. The same result can be seen from the 95% CI of (2.644; 11.130) which is greater than 1. The CI does not contain 1, hence there is a significant difference in the odds of an above average salary between data scientists who require python and those who do not.

Company Rating is a continuous predictor variable, since β_7 = 0.9128 > 0, this gives an $\widehat{OR}$ = 2.491 > 1, which indicates that the likelihood of a data scientist earning an above average salary increases as the company rating increases by one unit. Further, the 95% CI for the OR is given by (1.186; 5.233).

Resultantly, it can be concluded that Degree, Python and Company Rating are the factors that significantly contribute to a higher likelihood of a data scientist earning an above average salary.

7.2. The predicted probability

The first three observations are given as:

- Observation 1:

Salary:	Below average
Age:	48
Degree:	Undergraduate Degree
Company Rating:	3.8
Company Size:	501 – 1000
Founded:	1973
Type of Ownership:	Company – Private
Revenue:	\$5 to \$100 million (USD)
Python:	Yes

SAS:	Yes
Excel:	Yes
SQL:	No

- Observation 2:

Salary:	Below average
Age:	37
Degree:	Undergraduate Degree
Company Rating:	3.4
Company Size:	10000+
Founded:	1984
Type of Ownership:	Other
Revenue:	$2 to $10+ billion (USD)
Python:	Yes
SAS:	No
Excel:	No
SQL:	No

- Observation 3:

Salary:	Below average
Age:	56
Degree:	None
Company Rating:	3.8
Company Size:	1001 - 5000
Founded:	1965
Type of Ownership:	Government
Revenue:	$100 million to $2 billion (USD)
Python:	Yes
SAS:	No
Excel:	No

SQL: No

Fitting these into the reduced logistic regression model:

Observation 1: $\hat{\eta} = logit(\hat{y}) = -3.5724 + 1.8884(0) + 1.1784(1) - 2.2926(0) - 2.1127(1) - 0.9107(0) + 1.6910(1) + 0.9128(3.8)$

$= 0.65294$

$$\hat{\pi}(observation\ 1) = \frac{e^{\hat{\eta}}}{1 + e^{\hat{\eta}}} = \frac{e^{0.65294}}{1 + e^{0.65294}} = 0.6577$$

Observation 2: $\hat{\eta} = logit(\hat{y}) = -3.5724 + 1.8884(0) + 1.1784(1) - 2.2926(0) - 2.1127(0) - 0.9107(0) + 1.6910(1) + 0.9128(3.4)$

$= 2.40052$

$$\hat{\pi}(observation\ 2) = \frac{e^{\hat{\eta}}}{1 + e^{\hat{\eta}}} = \frac{e^{2.40052}}{1 + e^{2.40052}} = 0.9169$$

Observation 3: $\hat{\eta} = logit(\hat{y}) = -3.5724 + 1.8884(0) + 1.1784(0) - 2.2926(1) - 2.1127(0) - 0.9107(0) + 1.6910(1) + 0.9128(3.8)$

$= -0.70536$

$$\hat{\pi}(observation\ 3) = \frac{e^{\hat{\eta}}}{1 + e^{\hat{\eta}}} = \frac{e^{-0.70536}}{1 + e^{-0.70536}} = 0.3306$$

8. Conclusion

This report sought to identify the factors of the data scientist salary dataset which have a significant association with the salary of a data scientist. The tests of association concluded that the variables Degree, Python and SQL have a significant association with the response variable of interest, Salary. The reduced logistic regression model that was fitted to the data concluded that the factors Degree, Python and Company Rating significantly contribute to a higher likelihood of a data scientist earning an above average salary.

9. Appendix

9.1. SAS Code

```
/*CHANGING VARIABLES TO FOLLOW SAS NAMING CONVENTIONS*/
options validvarname=v7;

/* IMPORTING THE DATA */
proc import datafile = "/home/u58390843/STAT305/PROJECT/Data Scientist Salary Data.csv"
dbms=CSV
out= work.data_salary;
run;
proc contents data = work.data_salary;
run;

/* ONE-WAY TABLES FOR THE CATEGORICAL VARIABLES */
proc freq data=work.data_salary;
tables salary degree company_size type_of_ownership revenue python sas excel sql;
run;

/* SUMMARY MEASURES FOR THE QUANTITATIVE VARIABLES */
proc univariate data=work.data_salary;
var Age Company_Rating founded;
run;

/* SORTING THE ORDINAL DATA */
data work.salary_sorted;
set data_salary;
if salary="Below average" then salary="0_below average";
if salary="Above average" then salary="1_above average";
if degree="None" then degree="0_none";
```

```
if degree="Undergraduate Degree" then degree="1_undergrad degree";
if degree="Postgraduate Degree" then degree= "2_postgrad degree";
if revenue="Unknown / Non-Applicable" then revenue="0_revenue";
if revenue="$5 to $100 million (USD)" then revenue = "1_revenue";
if revenue="$100 million to $2 billion (USD)" then revenue= "2_revenue";
if revenue="$2 to $10+ billion (USD)" then revenue="3_revenue";
run;

/* TESTS OF ASSOCIATION */
proc freq data=work.salary_sorted;
tables degree*salary / chisq cmh ;
tables revenue*salary / chisq cmh ;
tables python*salary / chisq cmh;
tables sas*salary / chisq cmh;
tables excel*salary / chisq cmh;
tables sql*salary / chisq cmh;
run;

/* LOGISTIC REGRESSION MODEL */
proc logistic data=work.data_salary;
class salary degree (ref="None") revenue (ref="$2 to $10+ billion (USD)")
python (ref="No") sas (ref="No") excel (ref="No") sql (ref="No")/param=ref;
model salary = degree revenue python sas excel sql Age Company_Rating founded;
run;

/* REDUCED LOGISTIC REGRESSION, GOODNESS-OF-FIT TEST AND OVERDISPERSION TEST */
proc logistic data=work.data_salary;
class salary degree (ref="None") revenue (ref="$2 to $10+ billion (USD)") python (ref="No")
/param=ref;
```

```
model salary = degree revenue python Company_Rating / aggregate=(degree revenue python sas excel sql Age Company_Rating founded) scale=none lackfit;

run;


/*ACCOUNTING FOR OVERDISPERSION IN THE MODEL*/

proc genmod data=work.data_salary;

class salary degree (ref="None") revenue(ref="$2 to $10+ billion (USD)") python (ref="No") sas (ref="No") excel (ref="No") sql (ref="No");

model salary = degree revenue python Company_Rating/ aggregate=(degree revenue python sas excel sql Age Company_Rating founded) scale=pearson

dist=binomial link=logit;

run;
```

9.2. SAS Output

Appendix 1: One-way tables for the categorical variables

The FREQ Procedure

Salary	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Above average	94	44.34	94	44.34
Below average	118	55.66	212	100.00

Degree	Frequency	Percent	Cumulative Frequency	Cumulative Percent
None	104	49.06	104	49.06
Postgraduate Degree	31	14.62	135	63.68
Undergraduate Degree	77	36.32	212	100.00

Company_Size	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1 - 50	5	2.36	5	2.36
10000+	33	15.57	38	17.92
1001 - 5000	58	27.36	96	45.28
201 - 500	25	11.79	121	57.08
5001 - 10000	22	10.38	143	67.45
501 - 1000	56	26.42	199	93.87
51 - 200	13	6.13	212	100.00

Type_of_ownership	Frequency	Percent	Cumulative Frequency	Cumulative Percent
College / Univers	11	5.19	11	5.19
Company - Private	96	45.28	107	50.47
Company - Public	56	26.42	163	76.89
Government	12	5.66	175	82.55
Nonprofit Organiz	11	5.19	186	87.74
Other	26	12.26	212	100.00

Revenue	Frequency	Percent	Cumulative Frequency	Cumulative Percent
$100 million to $2 billion (USD)	79	37.26	79	37.26
$2 to $10+ billion (USD)	48	22.64	127	59.91
$5 to $100 million (USD)	44	20.75	171	80.66
Unknown / Non-Applicable	41	19.34	212	100.00

Python	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	117	55.19	117	55.19
Yes	95	44.81	212	100.00

sas	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	186	87.74	186	87.74
Yes	26	12.26	212	100.00

excel	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	96	45.28	96	45.28
Yes	116	54.72	212	100.00

sql	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	99	46.70	99	46.70
Yes	113	53.30	212	100.00

Appendix 2: Basic summary measures of the quantitative variables

The UNIVARIATE Procedure
Variable: Age

Moments			
N	212	Sum Weights	212
Mean	40.1556604	Sum Observations	8513
Std Deviation	11.2397111	Variance	126.331105
Skewness	0.2039857	Kurtosis	-1.2525509
Uncorrected SS	368501	Corrected SS	26655.8632
Coeff Variation	27.990353	Std Error Mean	0.77194653

Basic Statistical Measures			
Location		Variability	
Mean	40.15566	Std Deviation	11.23971
Median	39.00000	Variance	126.33111
Mode	25.00000	Range	35.00000
		Interquartile Range	21.50000

The UNIVARIATE Procedure
Variable: Company_Rating

Moments			
N	212	Sum Weights	212
Mean	3.62358491	Sum Observations	768.2
Std Deviation	0.5517833	Variance	0.30446481
Skewness	-0.1151191	Kurtosis	-0.2898273
Uncorrected SS	2847.88	Corrected SS	64.2420755
Coeff Variation	15.2275527	Std Error Mean	0.03789663

Basic Statistical Measures			
Location		Variability	
Mean	3.623585	Std Deviation	0.55178
Median	3.700000	Variance	0.30446
Mode	3.800000	Range	2.50000

The UNIVARIATE Procedure
Variable: Founded

Moments			
N	212	Sum Weights	212
Mean	1980.84434	Sum Observations	419939
Std Deviation	11.2397111	Variance	126.331105
Skewness	-0.2039857	Kurtosis	-1.2525509
Uncorrected SS	831860447	Corrected SS	26655.8632
Coeff Variation	0.56742021	Std Error Mean	0.77194653

Basic Statistical Measures			
Location		Variability	
Mean	1980.844	Std Deviation	11.23971
Median	1982.000	Variance	126.33111
Mode	1996.000	Range	35.00000
		Interquartile Range	21.50000

Appendix 3: Tests of association

Summary Statistics for Degree by Salary

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)				
Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	15.2282	<.0001
2	Row Mean Scores Differ	2	17.5394	0.0002
3	General Association	2	17.5394	0.0002

Summary Statistics for Revenue by Salary

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)				
Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	3.7741	0.0521
2	Row Mean Scores Differ	3	29.1476	<.0001
3	General Association	3	29.1476	<.0001

Summary Statistics for Python by Salary

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)				
Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	36.8150	<.0001
2	Row Mean Scores Differ	1	36.8150	<.0001
3	General Association	1	36.8150	<.0001

Summary Statistics for sas by Salary

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)				
Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	0.3829	0.5361
2	Row Mean Scores Differ	1	0.3829	0.5361
3	General Association	1	0.3829	0.5361

Summary Statistics for excel by Salary

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)				
Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	0.9053	0.3414
2	Row Mean Scores Differ	1	0.9053	0.3414
3	General Association	1	0.9053	0.3414

Summary Statistics for sql by Salary

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)				
Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	4.7651	0.0290
2	Row Mean Scores Differ	1	4.7651	0.0290
3	General Association	1	4.7651	0.0290

Appendix 4: The logistic regression model

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	293.172	226.462
SC	296.528	266.741
-2 Log L	291.172	202.462

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	88.7094	11	<.0001
Score	75.6051	11	<.0001
Wald	51.9863	11	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Degree	2	14.4088	0.0007
Revenue	3	20.2570	0.0002
Python	1	18.0168	<.0001
sas	1	0.0000	0.9958
excel	1	3.2291	0.0723
sql	1	0.7679	0.3809
Age	1	1.6414	0.2001
Company_Rating	1	4.4805	0.0343
Founded	0	.	.

Note: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

Founded =	2021 * Intercept - Age

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.2782	1.6439	1.9205	0.1658
Degree	Postgraduate Degree	1	1.9927	0.6136	10.5464	0.0012
Degree	Undergraduate Degree	1	1.2174	0.4163	8.5509	0.0035
Revenue	$100 million to $2 billion (USD)	1	-2.1238	0.5297	16.0786	<.0001
Revenue	$5 to $100 million (USD)	1	-2.0462	0.5814	12.3855	0.0004
Revenue	Unknown / Non-Applicable	1	-0.7535	0.5758	1.7124	0.1907
Python	Yes	1	1.6157	0.3806	18.0168	<.0001
sas	Yes	1	-0.00308	0.5875	0.0000	0.9958
excel	Yes	1	-0.6681	0.3718	3.2291	0.0723
sql	Yes	1	0.3583	0.4089	0.7679	0.3809
Age		1	-0.0217	0.0169	1.6414	0.2001
Company_Rating		1	0.8170	0.3860	4.4805	0.0343
Founded		0	0	.	.	.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Degree Postgraduate Degree vs None	7.335	2.204	24.419
Degree Undergraduate Degree vs None	3.378	1.494	7.640
Revenue $100 million to $2 billion (USD) vs $2 to $10+ billion (USD)	0.120	0.042	0.338
Revenue $5 to $100 million (USD) vs $2 to $10+ billion (USD)	0.129	0.041	0.404
Revenue Unknown / Non-Applicable vs $2 to $10+ billion (USD)	0.471	0.152	1.455
Python Yes vs No	5.031	2.386	10.609
sas Yes vs No	0.997	0.315	3.153
excel Yes vs No	0.513	0.247	1.062
sql Yes vs No	1.431	0.642	3.189
Age	0.979	0.947	1.012
Company_Rating	2.264	1.062	4.823

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	85.5	Somers' D	0.711
Percent Discordant	14.4	Gamma	0.711
Percent Tied	0.1	Tau-a	0.352
Pairs	11092	c	0.855

Appendix 5: The reduced logistic regression model

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	200.2852	113	1.7724	<.0001
Pearson	191.4115	113	1.6939	<.0001

Number of unique profiles: 121

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	293.172	223.923
SC	296.528	250.776
-2 Log L	291.172	207.923

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	83.2483	7	<.0001
Score	71.0572	7	<.0001
Wald	48.8034	7	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Degree	2	14.1898	0.0008
Revenue	3	23.7613	<.0001
Python	1	21.2679	<.0001
Company_Rating	1	5.8101	0.0159

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-3.5724	1.4323	6.2208	0.0126
Degree	Postgraduate Degree	1	1.8884	0.5823	10.5178	0.0012
Degree	Undergraduate Degree	1	1.1784	0.4008	8.6456	0.0033
Revenue	$100 million to $2 billion (USD)	1	-2.2926	0.5078	20.3867	<.0001
Revenue	$5 to $100 million (USD)	1	-2.1127	0.5740	13.5454	0.0002
Revenue	Unknown / Non-Applicable	1	-0.9107	0.5585	2.6595	0.1029
Python	Yes	1	1.6910	0.3667	21.2679	<.0001
Company_Rating		1	0.9128	0.3787	5.8101	0.0159

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	83.5	Somers' D	0.674
Percent Discordant	16.1	Gamma	0.677
Percent Tied	0.4	Tau-a	0.334
Pairs	11092	c	0.837

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
7.9455	8	0.4388

Appendix 6: Odds ratio estimates for the final model

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Degree Postgraduate Degree vs None	6.609	2.111	20.689
Degree Undergraduate Degree vs None	3.249	1.481	7.127
Revenue $100 million to $2 billion (USD) vs $2 to $10+ billion (USD)	0.101	0.037	0.273
Revenue $5 to $100 million (USD) vs $2 to $10+ billion (USD)	0.121	0.039	0.372
Revenue Unknown / Non-Applicable vs $2 to $10+ billion (USD)	0.402	0.135	1.202
Python Yes vs No	5.425	2.644	11.130
Company_Rating	2.491	1.186	5.233