**DATA401 Project**

**Critical Appraisal of Job satisfaction and wellbeing in New Zealand Stats**

**&**

**An Investigation into Graduate Admission Data**

**Part 1: Critical Appraisal of Job satisfaction and wellbeing in New Zealand Stats**

The job satisfaction and wellbeing from Stats NZ is a report which collect information from people to rate their satisfaction with their main job. The report was published on 23 July 2019 with the data collected by the Survey of Working Life 2018 (SoWL) and the New Zealand General Social Survey 2018 (GSS 2018) is in 2018. The main purpose of the report is to offer the insights of the job satisfaction in New Zealand. The data is mainly qualitative data and is presented using bar charts as the main mean of visualization. As there are many charts in the report, I intend to approach only 3 graphs of this report and focus on analysis of below 3 graphs:

+ Workplace relationship.

+ Overall life satisfaction rating by job satisfaction 2018.

+ Job satisfaction by mental wellbeing.

**Graph 1: Workplace relationship**

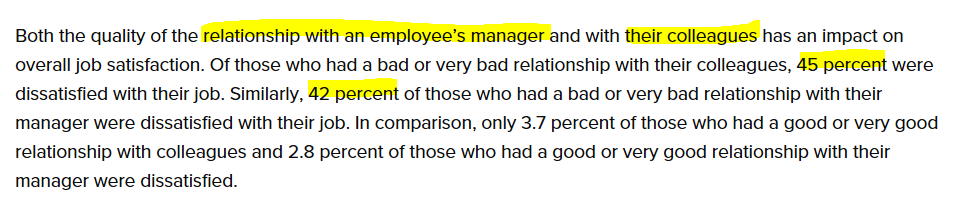


There are total four charts which show how each factor related to job satisfaction. This is one of those 4 charts which illustrate about how workplace relationship relate to job satisfaction. The data were qualitative data, so bar chart is a good approach to present the data. The format of this chart is clear as it didn’t use perspective graphs which usually appear in many reports as they distort the data.

The graph is supposed to illustrate those employees who have good overall workplace relationships that had much higher rates of satisfaction (91 percent) than those who did not (55 percent). The number illustrate in the report was correct. However, they rounded the percentage without stating it was rounded to the whole number (the actual number on chart is 55.2% which rounded in the report to 55%).

There is an analysis from the report which put below the graph:

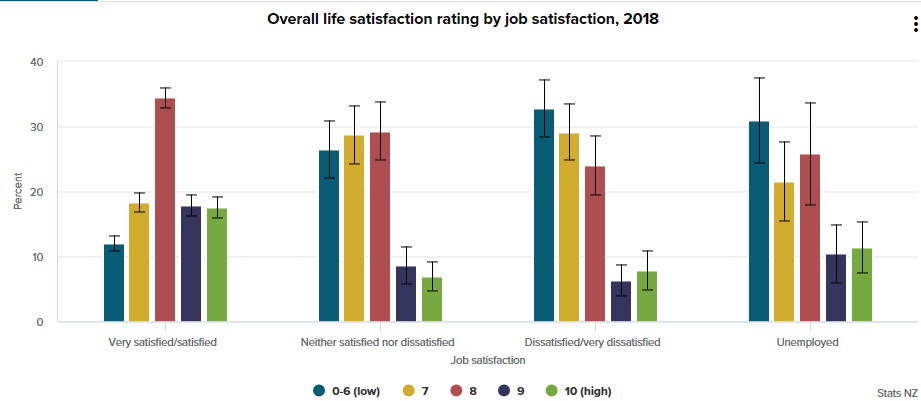
***Figure 1: The comment of the graph 1 – extracted from the report***

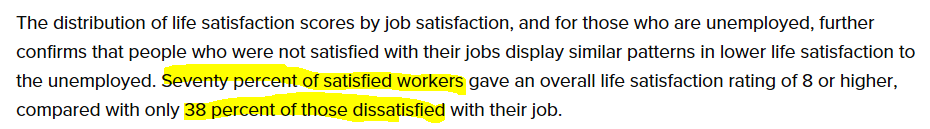


When reading this data analysis, what make audience confused at the first glance is: In the graph, we couldn’t see any information about relationship with an employee’s manager and with their colleague and these distinguished relationship relate to job satisfaction. The graph just shows the workplace relationship as a whole (good relationship –not good relationship). It’s not easy to spot these kind of numbers (45 percent, 42 percent).

However, set aside from the unclear data, the graph was already successfully conveying the idea of how the workplace relationship relate to job satisfaction with easily understood manner for audiences.

**Graph 2: Overall life satisfaction rating by job satisfaction 2018.**





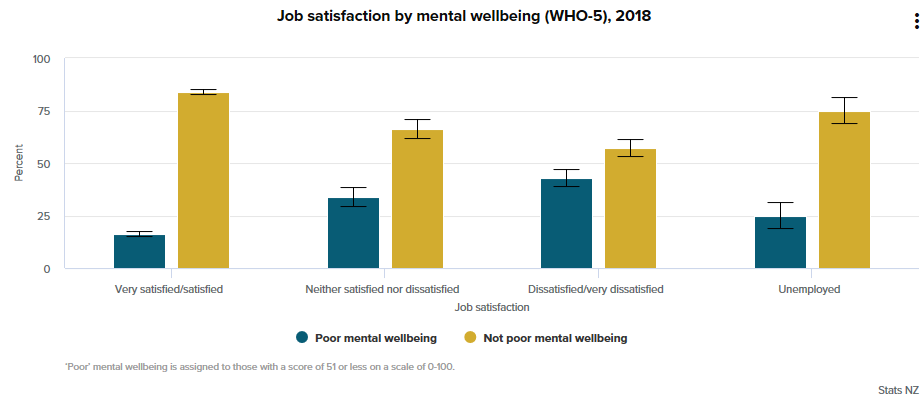
Source: <https://www.stats.govt.nz/reports/job-satisfaction-and-wellbeing>

The bar chart illustrates the distribution of life satisfaction scores by job satisfaction. This is qualitative data (ordinal data), so the bar chart is a good choice to display this kind of data.

What could be confusing is the comment for this chart. “Seventy percent of satisfied worker” as compared to “38 percent of those dissatisfied with their job”, the total percentage here is **over** 100 percent already (70% + 38% = 108%), which should be only 100% as a normal rule of maximum percentage.

Set aside this inaccurately drawn graph, the chart clearly shows the satisfied workers gave very high rating (8 or above) while dissatisfied workers and unemployed workers gave very low rating for life satisfaction (7 or below), which means the job satisfaction have impact on life satisfaction.

**Graph 3: Job satisfaction by mental wellbeing**



This chart illustrate “only 16 percent of those who were satisfied with their main job were classified as having poor mental wellbeing. This increased to 43 percent of those who were dissatisfied with their job. This was significantly higher than the proportion of the unemployed population with ‘poor’ mental wellbeing (25 percent).”

The data was collected based on the answer of respondents, who required to answer 5 different questions which derived from a set of mental wellbeing questions developed by the World Health Organization (WHO).

The error bars here are graphical representations of the variability of data and used on graphs to indicate the error or uncertainty in a reported measurement.

The only limitation of this chart is: we can’t see the percentage between satisfied and dissatisfied workers as a whole as the data has been grouped into 4 independent groups with percentage of each group equal to 100%. We also couldn’t view how large the gap between the percentage of poor mental wellbeing as compared to not poor mental wellbeing is. Therefore, it would be helpful if this StatsNZ report show one chart (it can be pie chart or bar chart) to see how many percentage for each case (poor mental wellbeing vs not poor mental wellbeing) first, so that audience can have an overview about these category before transferring to the “Job satisfaction by mental wellbeing” graph.

**Alternative methods**

A number of alternative methods could have been used to present this data:

1. Most of the data is qualitative data. It would be helpful if this report includes the quantitative factor such as the earning vs Job satisfaction which maybe of curiosity of audiences.
2. The data which StatsNZ collected is in 2018. The report was published on 23 July 2019 with bar-charts as the main mean of visualization of the data. Time series plots could have been used to show the trend of job satisfaction over time for audiences to see if there is a reduction or an increase in job satisfaction over time. This would be the very helpful information for audiences.
3. In the report, StatsNZ stated: “We looked at some of the characteristics of people’s jobs, such as hours and times of work, job security, workplace autonomy and relationships, and work-related stress and tiredness, to see how they each relate to job satisfaction.” The question is why other factors couldn’t be included in this report. What is the main reference resources for them to choose these 5 main factors? The OECD has developed a framework to measure and assess the quality of jobs that considers three objective:

**+ Earnings quality** captures the extent to which earnings contribute to workers' well-being in terms of average earnings and their distribution across the workforce.

**+ Labour market security** captures those aspects of economic security related to the risks of job loss and its economic cost for workers. It is defined by the risks of unemployment and benefits received in case of unemployment.

**+ Quality of the working environment** captures non-economic aspects of jobs including the nature and content of the work performed, working-time arrangements and workplace relationships. These are measured as incidence of job strain characterised as high job demands with low job resources.

We can see that the StatsNZ has presented the main factors which related to job satisfaction such as job security (belongs to Labour market security of OECD); hours and times of work, workplace autonomy and relationships, and work-related stress and tiredness (belongs to Quality of the working environment category of OECD). However, it would be helpful if they can state clearly the reason why they choose these 5 factors as there are still some of factors which they miss mentioning like: Earning quality, nature and content of the work performed.

Another improvement which I think it would be better if StatsNZ could add on self-employment segment in their report. In this report, StatsNZ mainly explore the job satisfaction status of paid-employment segment. Actually, we still have another kind of employment, self-employment, which also need to be explored to see which segment have high job satisfaction than the other. It would be great if they show the chart of percentage between paid-employment and self-employment to shows evidence that paid-employment segment constitutes the extreme large proportion of employment in New Zealand, therefore this segment is deserved to do research than self-employment segment.

**Conclusion**

The report communicates job satisfaction data in an easy to understand format for the intended audience, the general public. Overall, this report has been successful in delivering data insight to the audience with a straight forward and clear way. However, this report could be improved more by stating clearly why they choose paid-employment segment to explore and the reason why they choose these kind of factors to make statistic comparison. The data also would be better expressed if time series graph were included to show the trend of job satisfaction overtime.

**Part 2: An Investigation into Graduate Admission Data**

**Introduction**

This dataset is created for prediction of Graduate Admissions from an Indian perspective. This dataset is inspired by the University of California, Los Angeles (UCLA) Graduate Dataset. The dataset is owned by Mohan S Acharya. The dataset contains several parameters which are considered important during the application for Masters Programs. The parameters included are:

1. The Graduate Record Examinations (GRE) Scores (out of 340).

2. The Test of English as a Foreign Language (TOEFL) Scores (out of 120).

3. University Rating (out of 5).

4. Statement of Purpose and Letter of Recommendation Strength (out of 5).

5. Undergraduate Grade Point Average (GPA) (out of 10).

6. Research Experience (either 0 or 1).

7. Chance of Admission (ranging from 0 to 1).

Chance of Admission will be impacted by all of the predictors (GRE Scores, TOEFL Scores, Statement of Purpose and Letter of Recommendation Strength, Undergraduate GPA, Research Experience). However, based on the scope of the Project and knowledge which I gained from the DATA401 course, I intend to use this data set for figure out if there is the relationship between GRE score, TOEFL score with the chance of admit.

Firstly, we use Summary Statistics to summarise data, produce graphs.

Figure 1: GRE score of students who applied for Master Programs at UCLA in 2019.

Figure 2: TOEFL score of students who applied for Master Programs at UCLA in 2019.

Summary about descriptive statistics about GRE score and TOEFL score of students who applied for Master Programs at UCLA in 2019 as below table:

|  |  |  |
| --- | --- | --- |
|  | **GRE Score** | **TOEFL Score** |
| Mean | 316.472 | 107.192 |
| Median | 317 | 107 |
| Standard Deviation | 11.295 (3d.p) | 6.082 (3d.p) |

Looking at the distribution of GRE score of students who applied for Master Programs at UCLA in 2019, we see an almost bell shaped distribution with the mean of 316.472 (3dp) and median of 317.000 (3dp), so data are little bit skewed to the left (**not heavily skewed**). **The sample size is big** (n>30), so based on Central Limit Theorem, we can approximate this distribution as **Normal distribution**.

Similarly, we can see that the distribution of TOEFL score of students who applied for Master Programs at UCLA in 2019 is also bell shaped with the mean of 107.192 (3dp) and median of 107.000 (3dp), so skewed to the left.

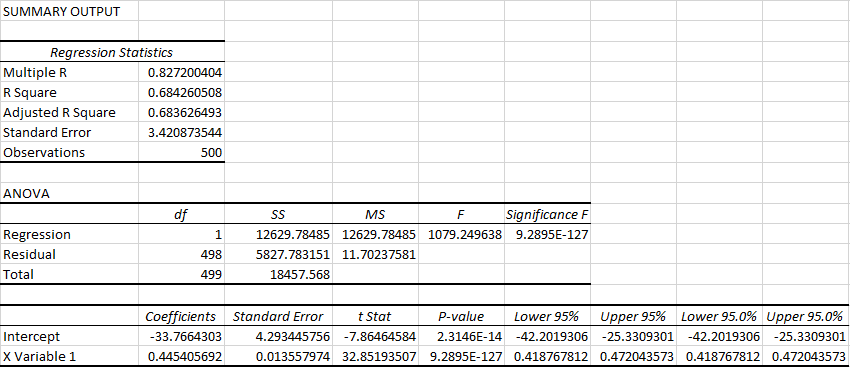
*Figure 3: TOEFL score and GRE score of students who applied for Master’s program at UCLA 2019.*

On one hand, when GRE scores and TOEFL scores of students who applied for Master’s program at UCLA 2019 are plotted on the scatter plot, they appear to have a linear relationship. The student who have high GRE score are likely to have high TOEFL score too.

TOEFL score = 0.4454\* (GRE score) – 33.766. The model indicates that for every mark increase in GRE score, a student is likely to increase on average 0.445 mark (3dp) more in TOEFL scores. The y-intercept in this model of -33.766 (3dp) does not make a lot of sense in this context as a student cannot have 0 score in GRE score, nor can they have a negative score, so it is not useful for extrapolation in this situation.

Looking at the regression output below we can see that the Multiple R value (coefficient of correlation) is 0.827 (3dp) which indicates that there is a high positive correlation between TOEFL score and GRE score. The R-square (coefficient of determination) value shows that 68.4% (1dp) of variation in TOEFL score can be attributed to variation in GRE score.

On the other hand, if we choose GRE score as response variable instead of TOEFL score to investigate their relationship, we have GRE score = 1.54\* (TOEFL score) + 151.80 (2dp) with the same Multiple R and R-square value (see Appendix for scatter plot and Regression output for choosing y = GRE score). This means that GRE score and TOEFL score, no matter which is the predictor and which is response variable, their relationship between two variables in 2 cases are the same (strong and positive relationship). So in this report, I have just focused in one case (y = TOEFL score) to do Inferential Statistics.



A Hypothesis test of the slope will determine if there is evidence for a relationship between TOEFL scores and GRE scores of student who applied Master’s program at UCLA 2019.

**Hypothesis Test the Slope Parameter** 𝜷𝟏

𝐻0: 𝛽1 = 0

𝐻𝑎: 𝛽1 ≠ 0

𝛼 = 0.05

df = 500 – 2 = 498

Test Statistic: 𝑇𝑎/2, n−2 = 32.85 (2𝑑𝑝).(𝐹𝑟𝑜𝑚 𝑅𝑒𝑔𝑟𝑒𝑠𝑠𝑖𝑜𝑛 𝑂𝑢𝑡𝑝𝑢𝑡)

Critical Value: t critical = T.INV.2T (𝛼, df) = T.INV.2T (0.05, 498) = ± 1.965 (3dp)

Decision: Reject H0 because 32.85 > 1.965

Conclusion: There is strong evidence at α = 0.05 to suggest that the linear model is useful for predictions. The GRE score is a useful predictor for TOEFL score.

The 95% confidence interval for β1 is (0.419, 0.472) (3dp.) (from Regression Output)

We are 95% confident that the true population slope lies in the interval (0.419, 0.472). The number 0 is not contained within the confidence interval and so 𝐻0 can be rejected at a 95% confidence level.

**Chance of Admission of students**

Chance of admission (ranging from 0 to 1) represents the probability of a student who successfully applied to Master’s Program at UCLA (University of California, Los Angeles, USA). The chance of admission will be higher if the number is closer to 1. Figure 4 illustrates the distribution of chance of admission of students who applied Master’s program at UCLA in 2019.

*Figure 4: Chances of admission of students who applied Master‘s program at UCLA in 2019*

Descriptive Statistics Table:

|  |  |
| --- | --- |
| Mean | 0.72174 |
| Median | 0.72 |
| Standard Deviation | 0.141140404 |

Looking at the distribution of chance of admission of students who applied for Master Programs at UCLA in 2019, we see an almost bell shaped with the mean of 0.722 (3dp) and median of 0.72 (2dp), so data are little bit skewed to the right (not heavily skewed). The sample size is big (n>30), so based on Central Limit Theorem, we can approximate this distribution as Normal distribution.

We do not have the entire population of all students who applied Master‘s program at UCLA 2019. The chance of admission have mean of 0.722 (3dp). Our research question is: **Is it true that more than 50% of students have chance of admission greater than 0.7?**

This question is deserved to evaluate and build testing hypothesis because it can help us to make inferences or predictions about the proportion of the population who have high chance of admission (if chance of admission is greater than 0.7, this means that their applications are more likely to be approved by UCLA).

For this question, we defines “the number of successes” will be the number of students who have chance of admission is greater than 0.7.

State the question statistically in terms of population parameters: p > 0.5

State the complement statistically: p ≤ 0.5

Now write the hypothesis:

𝐻0: p = 0 .5

𝐻𝑎: p > 0.5

Check requirements:

np0 = 500 \* 0.5 = 250 > 5;

nq0 = 500 \* 0.5 = 250 > 5

🡪 satisfy the requirement that: the Normal approximation can be used

==

= 0.574 (3dp)

**Hypothesis Test the proportion:**

𝐻0: p = 0 .5

𝐻𝑎: p > 0.5

𝛼 = 0.05

n = 500

**Test Statistic:** z = = = 3.309 (3dp)

**Critical value:** z = NORM.S.INV(0.95) = 1.645 (3dp) (right tail)

**Decision:** Reject H0 because 3.309 > 1.645

**Conclusion:** There is sufficient evidence at 𝛼 = 0.05 to suggest that: there is more than 50% of students have chance of admission greater than 0.7

**90% Confidence Interval:**

1.645 \* = 0.574 1.645 \* = (0.538; 0.610) (3dp)

The 90% confidence interval for the proportion of students who have chance of admission greater than 0.7 is (53.76%, 61.03%) (2dp)

This means we are 90% confident that the true proportion of students who have chance of admission greater than 0.7 lies in the interval (53.76%, 61.03%). The number 50% (0.5) is not contained within the confidence interval and so 𝐻0 can be rejected at a 90% confidence level.

**The relationship between GRE score and TOEFL score with Chance of Admission of students**

Chance of Admission will be impacted by two predictors: GRE score and TOEFL score. However, to check how strong this relationship is and to test if we have sufficient evidence for a relationship between scores (TOEFL scores and GRE scores) and chance of admission, we need to build hypothesis to test it.

In section 1, we have concluded that GRE score and TOEFL score have positive correlation. The students who have high GRE score are likely to have high TOEFL score and vice versa. In this section, to reduce the complexity when exploring the relationship between scores and chance of admission, we assume that GRE score and TOEFL score have the same relationship with chance of admission. Then, it is relevant to use one kind of score (e.g. GRE score) as a representative to test the relationship with chance of admission.

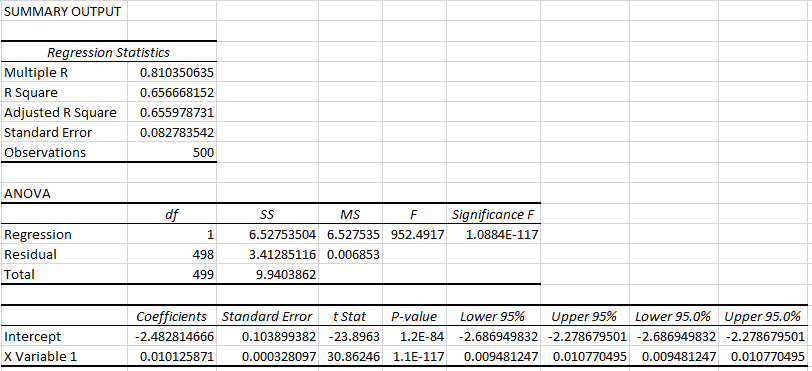
First, we plot the scatter plot to investigate the relationship.

*Figure 5: GRE score and chance of admission of students who applied for Master’s program at UCLA 2019*

When GRE score and chance of admission of students who applied for Master’s program at UCLA 2019 are plotted on the scatter plot, they appear to have a linear relationship. The student who have high GRE score are likely to have high chance in admission.

Chance of admission = 0.010\* (GRE score) – 2.483 (3dp). The model indicates that for every mark increase in GRE score, a student is likely to increase on average 0.010 (~ 1%) chance of admission (3dp). The y-intercept in this model of -2.483 (3dp) does not make a lot of sense in this context as a student cannot have 0 score in GRE score, nor can they have a negative score, so it is not useful for extrapolation in this situation.

Looking at the regression output below, we can see that the Multiple R value (coefficient of correlation) is 0.810 (3dp) which indicates that there is a high positive correlation between GRE score and chance of admission. The R-square (coefficient of determination) value shows that 65.7% (1dp) of the variation in Chance of Admission can be explained by the variation in GRE score.



A Hypothesis test of the slope will determine if linear model are useful for prediction in this case or not.

**Hypothesis Test the Slope Parameter 𝜷𝟏 between GRE score and chance of admission**

𝐻0: 𝛽1 = 0

𝐻𝑎: 𝛽1 ≠ 0

𝛼 = 0.05

df = 500 – 2 = 498

**Test Statistic:** 𝑇𝑎/2, n−2 = 30.86 (2𝑑𝑝). (𝐹𝑟𝑜𝑚 𝑅𝑒𝑔𝑟𝑒𝑠𝑠𝑖𝑜𝑛 𝑂𝑢𝑡𝑝𝑢𝑡)

**Critical Value**: t critical = T.INV.2T (𝛼, df) = T.INV.2T (0.05, 498) = ± 1.965 (3dp)

**Decision**: Reject H0 because 30.86 > 1.965

**Conclusion:** There is strong evidence at α = 0.05 to suggest that the linear model is useful for predictions. The GRE score is a useful predictor for chance of admission.

The 95% confidence interval for β1 is (0.009, 0.011) (3dp.) (from Regression Output)

We are 95% confident that the true population slope lies in the interval (0.009, 0.011). The number 0 is not contained within the confidence interval and so 𝐻0 can be rejected at a 95% confidence level.

**Conclusion**

When investigating Graduate Admission dataset, we can see that the GRE score and TOEFL score have positive correlation. The students who have high GRE score are likely to have high TOEFL score and vice versa. The hypothesis test for the slope also shows strong evidence that the GRE score is a useful predictor for TOEFL score and we can use linear model for prediction.

The chance of admission represents the probability of a student who successfully applied to Master’s Program at UCLA (University of California, Los Angeles, USA) and in our hypothesis test, there is sufficient evidence at 𝛼 = 0.05 that over 50% of students who have chance of admission greater than 0.7. This means that in our dataset, over 50% of students have high chance to receive the acceptance in their application to Master’s Program at UCLA in 2019.

About the relationship between score and chance of admission, we can see that high positive correlation between GRE score and chance of admission. After using hypothesis test to test the relationship between GRE score and chance of admission, we have strong evidence that we can use GRE score as a useful predictor of chance of admission of students.

**References**

Stats NZ Tatauranga Aotearoa. (2019, July 23). *Report about Job satisfaction and wellbeing.*

Retrieved from <https://www.stats.govt.nz/reports/job-satisfaction-and-wellbeing>

OECD - Organisation for Economic Co-operation and Development*. Job quality*.

Retrieved from <https://www.oecd.org/statistics/job-quality.htm>

Mohan S Acharya, Asfia Armaan, Aneeta S Antony : *A Comparison of Regression Models for Prediction of Graduate Admissions*, IEEE International Conference on Computational Intelligence in Data Science 2019

Retrieved from <https://www.kaggle.com/mohansacharya/graduate-admissions>

**Appendix**

GRE score is the response variable

