The mirage of SETs: Do teaching scores predict student salaries?

Kiss Viktor
MSU Denver

Edgar Maldonado
MSU Denver

Vicky Seehusen
MSU Denver

ABSTRACT

This paper presents a study that analyzed the relationship between Student Evaluation of Teaching (SET) and mid-career median pay and was conducted at a state university. The analysis was based on 49 courses prefixes from 5 different colleges/schools within the university. The SET data was collected from a wide range of courses, and the analysis revealed a negative correlation between SET and earning, indicating that courses with lower ratings for instruction are associated with higher mid-career median pay. A further analysis of the relationship found the quantitative character of the discipline as a confounding variable. The relation became clear: lower SET scores are correlated with quantitative majors, while quantitative majors are correlated with higher salaries. These findings have implications for how universities and colleges evaluate and reward instructors, as well as how students perceive and value their educational experiences.

Keywords: SET, Salary, Quantitative, Qualitative, Discipline

INTRODUCTION

In a world where the rating of products and services are rituals for the general consumer and the use of such ratings are embedded in most purchasing activities, it seems intuitive to use Student Evaluation of Teaching (SET) as a tool to rate the quality of Education. Additionally, a quantitative measure that is often used associates salary with education level; according to the U.S. Bureau of Labor Statistics (2018) “the more you learn, the more you earn”. Thus, SET and salary seem to be good candidates for exploring a relationship between how favorably a student considers a professor's teaching and the expected salary of that student. This research aims to delve into this relationship and shed light on the role of an additional variable that may introduce what is commonly referred to as a ‘spurious correlation’, thereby providing valuable insights into the dynamics at play in educational and economic outcomes.

STUDENT EVALUATION OF TEACHING

The Student Evaluation of Teaching (SET) is a tool used to measure the quality of the teaching in higher education. The procedure consists of collecting students’ opinions about teaching aspects of a faculty. Students’ opinions are translated into numerical values that are used to (1) improve teaching and learning, (2) inform managerial decisions related to educators’ job security, (3) inform prospective students’ decisions on choosing academic programs, and (4) comply with governments’ requirements (Palmer, 2012).

Although SETs are a staple of most academic institutions, they have been a source of debate since their introduction. The validity of SET instruments has been widely questioned in the literature (Uttl, 2021). The arguments against the artifact can be organized in two categories (Ali et al., 2021): validity and reliability. In the validity realm of the arguments, authors dispute how adequate the tool is to measure efficient teaching. For example, Clayson (2022) argues how SETs fail to capture the multidimensionality aspect of teaching, and Uttl (2021) points to how the definition of good teaching is not clear yet. Additionally, some authors question the reliability of the tool. For example, Ali et al. (2021) explains how factors unrelated to faculty teaching, such as personal likability of the teacher/professor, play a critical role in SETs. Uttl includes gender, ethnicity, and the attractiveness of the teacher/professor in the list of unrelated factors that play a role in SETs (2021).

Despite generating some debate, Student Evaluation of Teaching is still one of the main resources for university administrators in charge of faculty promotions and merit increases (Stroebe, 2020).

SALARIES AND EDUCATION

There are other ways to measure education effectiveness. Sokoli (2020) summarizes the Organization for Economic Cooperation and Development (OECD) framework to measure education using 4 categories: context, resources, processes, and results. The context dimension is related to the social and economic background of the educational environment. The resources and processes are related to “what and how” education takes place. Finally, results are related to finishing degrees, employment, and *salaries*.

According to Paré (2023), salaries and career outcomes are also used by some higher education institutions as a marketing tool for recruiting. Therefore, the idea of salaries as a dimension to measure quality of education is not a novel one. Since SETs are a tool to also measure the quality of education, it seems intuitive to study the relation between SETs and salaries.

JUSTIFICATION

The use of correlation and regression methods to study SETs is common in academic literature. For example, Esarey and Valdes (2020) explain how the SETs have been used to find correlations with other measurements such as Students' Learning and Teaching Quality. The authors of this research take a different approach by looking for a common measure of Education that does not take place in academic settings: salaries.

Although the original premise for this study came from the notion of analyzing the relationship between SETs and salaries, a negative correlation found in preliminary analysis prompted the authors to explore the use of additional variables to explain the relationship.

METHODOLOGY

Our analysis is grounded in data collected from mid-size urban state university in the Rocky Mountains with a student population of around 23,000 students. The data was collected from 2012 to 2022 (inclusive), spanning a period of eleven years. Specifically, the data provided for this study includes information on STEs from 76793 sections and 3572 courses, for a total of 748746 student responses. The SETs in this institution uses a 1-6 simple scale. To draw meaningful conclusions from the data, this study only considers prefixes with at least 3000 student ratings, leaving 51 prefixes to analyze. A larger sample size is more likely to provide a representative view of the student population's opinions about the courses and increase the statistical power of the study. Because of this, the data provides valuable insights into the student experience and perceptions of the courses included in the study. Please find the full list of prefixes in Appendix A.

For the salary data, the research uses information collected by PayScale, Inc. a company that provides information and tools related to compensation and salary data. The company reports the median Mid-Career Pay (10+ years of experience) for a wide variety of majors, and its data is based on the response of 3.5 million respondents across the United States (https://www.payscale.com/college-salary-report).

The following plot contains all of the prefixes plotted by their SETs and salaries. The full list of prefix abbreviations can be found in Table 3 (Appendix). The first step of the research is to establish whether SET and salaries are associated. A scatterplot of the data is shown in Figure 1 (Appendix). Since there is only one explanatory variable in the model, the choices for numerical methods for analysis are limited. Due to the simplicity and relatively high interpretability, the authors chose to explore different degrees of polynomial regression to see which one fit the best. Once fitted, the Bayesian Information Criterion (BIC) was used to choose the best fitting degree. BIC is a way to balance how well a model fits the data with how complex it is. It rewards models that fit the data well but are not too complex and penalizes models that are too complex. The results showed that the first-degree polynomial (a regression line) had the lowest BIC and thus would be used for analysis. Homoskedasticity and the normality of residuals were checked before obtaining the results of the model. Figure 2 shows the regression line, which yielded the following results:

The model is highly significant (p-value: 0.0000077), showing a lower salary by $70,000 for each unit drop in SET. This low p-value indicates that the relationship between Student Evaluation of Teaching scores and mid-career pay is highly unlikely to have occurred by random chance. The correlation coefficient between SET and mid-career pay is 0.58, meaning as SET scores increase, mid-career pay tends to decrease as well, and vice versa. This correlation coefficient falling between 0.5 and 0.7 indicates a significant positive association between the two variables.

The standard error is moderate: $14,887 which is around 19% of the range of the salaries. A moderate standard error like this suggests that the predicted mid-career pay might have some variance around the model's predictions, but still makes for a robust model.

The R-squared value of 33.8% adds another layer of insight. Although not accounting for the entire variability, the model does a moderately good job at explaining more than 30% of the fluctuations in mid-career pay using SET scores.

The presented statistical analysis showcases a strong case for a negative relationship between SET scores and mid-career pay. It is important to note that the authors in no way imply that there is a direct *causal* relationship between SETs and Salary. Since mid-career wages are not *directly* influenced by the average SET the course prefixes got in a given student's major, the relationship seems to be spurious, suggesting that there might be confounding variable(s) that effect both SET and Salary. This leads us into our next section.

Causal framework

As correlation between variables does not necessarily mean a causal relationship, it is worth exploring whether indications of causality can be found within the model. A simple Directed Acyclic Graph. Directed Acyclic Graphs (DAGs) serve as the backbone of causal reasoning by providing a visual and formal framework to depict causal relationships among variables. In a DAG, nodes represent variables, while arrows represent casual connections in the direction of the arrow. The structure of DAGs in causal reasoning does not come from the data, but from domain knowledge. The data serves as a way to disprove the validity of a hypothesized DAG. If one cannot refute a DAG with the data, it means that the data could have been generated by that causal structure, with the possibility that that there is another DAG within the same equivalence class, since DAGs within the same equivalence class produce identical implications for conditional independence relationships among variables.

Figure 3 (Appendix) is a plausible causal diagram which would imply that in the presence of confounders, there would be an observable relationship between SETs and Salary. If a new model could control for all such confounder(s), the relationship between SETs and Salary should in principle go away.

One such potential confounder that the authors have found in the literature is whether the disciplines rely on quantitative methods or not. Uttl & Smibert's (2017) performed a study with more than 14,000 SETs and found a statistical difference between the average SET of a quantitative course against a non-quantitative one. The authors explored this venue by first categorizing the majors in the study as quantitative or non-quantitative. The authors used University descriptions of disciplines and also turned to the Department of Homeland Security STEM Designated Degree Program List (2023) as a guideline to classify the majors (Table 3, Appendix). The working hypothesis is that the quantitative nature of a discipline affects the SETs and also the ultimate salary as seen in the new DAG below.
In this case, if a new model control for the quantitative variable (named quant), the relationship between SETs and Salary should vanish, depending on the influence of the remaining other confounders which are not controlled in the analysis.

For this, the dataset was divided into two sections: quantitative vs non-quantitative disciplines, effectively blocking confounding effect of the Quantitative Majors variable. A separate model was run for each of the datasets; the results can be seen below. For both quantitative and non-quantitative disciplines, the relationship between SET scores and salary appears to have diminished considerably. The p-values for both regression models (23.85% and 10.83%) indicate that these relationships are not statistically significant at the conventional 5% level, meaning there is no reliable association left between the variables. Also, the explanatory power of SET on the variation of salaries has dropped to 4.6% and 14.46% (R^2) for non-quant and quant prefixes respectively. This is very low compared to the previous 33.8%. The drastic decrease in the strength of the relationship between SET scores and salary, coupled with non-significant p-values and minimal explanatory power, suggests that SET scores might not be as predictive of salary once controlling for the quantitative nature of the disciplines.

These results imply that the DAG in Figure 5 could have plausibly generated the data, since the association between SET and Salary almost completely vanished. The p-values indicate that it is most likely that the original model did not account for all the confounding, and there might still be relatively minor lingering confounding factors, but the vast majority of the association was accounted for with the new model.

Validating the effect of the Quantitative variable on SET and Salary

In order to confirm that the relationships between the confounding variable (Quantitative variable) and SET and Salary are valid, the standard Welch's t-test (Welch, 1947) was individually conducted for each of them. To validate the assumptions, Shapiro-Wilk tests were conducted to evaluate the normality of the data for Quant = 0 and Quant = 1 in both SET and Salary. The tests yielded high p-values, suggesting that the data for both groups does not significantly deviate from a normal distribution. The t-tests were conducted, which showed that for both SET and Salary there is a highly significant difference between Quantitative and non-Quantitative disciplines. The difference in SET average scores is 0.18, while the Salary difference amounts to -$26,760. Please see Table 4, Appendix for details.

This suggests a systematic difference in SETs based on the quantitative nature of the discipline, as visually demonstrated in Figure 6. This difference in SETs and Salaries based on the quantitative nature of the discipline provides evidence that the quantitative nature can indeed be a confounding variable between SET and Salary further validating the hypothesized DAG in Figure 4.

IMPLICATIONS

We can see that by conditioning on the Quantitative variable, the new model managed to block the vast majority of the association between SETs and Salary, meaning that the correlation was highly driven by whether the disciplines were quantitative in nature rather than any direct relationship. These findings hold implications for the whole academic realm. While the initial negative correlation could lead to instructors feeling that lower SETs are “justified” since it will lead to higher salaries, that is not likely the case. The pronounced shift in correlation when accounting for quantitative nature of the discipline emphasizes that SETs might not necessarily be indicative of instructors’ contributions to student success and future earnings.

Academic institutions should also consider adopting discipline-specific evaluation metrics for assessing teaching effectiveness. This approach ensures that instructors are evaluated based on criteria relevant to their specific fields, acknowledging the diversity of teaching methods and learning outcomes across disciplines.

CONCLUSION AND FUTURE RESEARCH

In this research the authors have explored the relationship between SETs and salaries. Both measurements have been used before as indicators of education quality, but the relation between them was unexplored. When investigating this relationship using majors as unit of analysis, our analysis indicated that a negative relationship between the two magnitudes is the result of a confounding variable: the quantitative character of the major.

According to the results, Student Evaluation of Teaching are not good predictors of salaries. Nevertheless, SETs seem to have a different impact when considering the quantitative character of the majors. The fact that SETs could be correlated with salaries could be considered an indicator of the complexity of evaluating education. Therefore, institutions and students must take care not to evaluate faculty by looking exclusively at the SETs.

Future research could study SETs within specific fields to develop a more precise tool for analyzing the performance of professors. An initial categorization of quantitative and non-quantitative courses could be a starting point for developing a fairer use of SETs in faculty evaluation.

REFERENCES

Ali, A., Crawford, J., Cejnar, L., Harman, K., & Sim, K. N. (2021). What student evaluations are not: scholarship of Teaching and Learning using student evaluations. *Journal of University Teaching & Learning Practice*, 18(8), 01.

Clayson, D. (2022). The student evaluation of teaching and likability: what the evaluations actually measure. *Assessment & Evaluation in Higher Education*, 47(2), 313-326.

Department of Homeland Security (2023). DHS STEM Designated Degree Program List. Last accessed 9/01/2023. https://www.ice.gov/doclib/sevis/pdf/stemList2023.pdf

Esarey, J., & Valdes, N. (2020). Unbiased, reliable, and valid student evaluations can still be unfair. *Assessment & Evaluation in Higher Education*, 45(8), 1106-1120.

Palmer, S. (2012). Student evaluation of teaching: keeping in touch with reality. *Quality in higher education*, 18(3), 297-311.

Paré, R. (2023). The limits of Salary as Measure of University Success. The evolution. Last accessed in 8/26/2023. https://evolllution.com/technology/metrics/the-limits-of-salary-as-a-measure-of-university-success/

Sokoli, D. (2020). Indicators of quality in Higher Education – Literature review. *UBT International Conference*. 109.

Stroebe, W. (2020). Student evaluations of teaching encourages poor teaching and contributes to grade inflation: A theoretical and empirical analysis. *Basic and Applied Social Psychology*, 42(4), 276-294.

U.S. Bureau of Labor Statistics (2018). Measuring the value of education. Last accessed 9/1/2023: https://www.bls.gov/careeroutlook/2018/data-on-display/education-pays.htm.

Uttl, B. (2021). Lessons learned from research on student evaluation of teaching in higher education. Student Feedback on Teaching in Schools: Using Student Perceptions for the Development of Teaching and Teachers, 237-256.

Uttl, B., & Smibert, D. (2017). Student evaluations of teaching: teaching quantitative courses can be hazardous to one's career. *PeerJ*, 5, e3299.

Welch, B. L. (1947). The generalization of "Student's" problem when several different population variances are involved. *Biometrika*. 34 (1–2): 28–35.

APPENDIX

Figure 1 - Scatter Plot of Set and Mid-Career Median Income

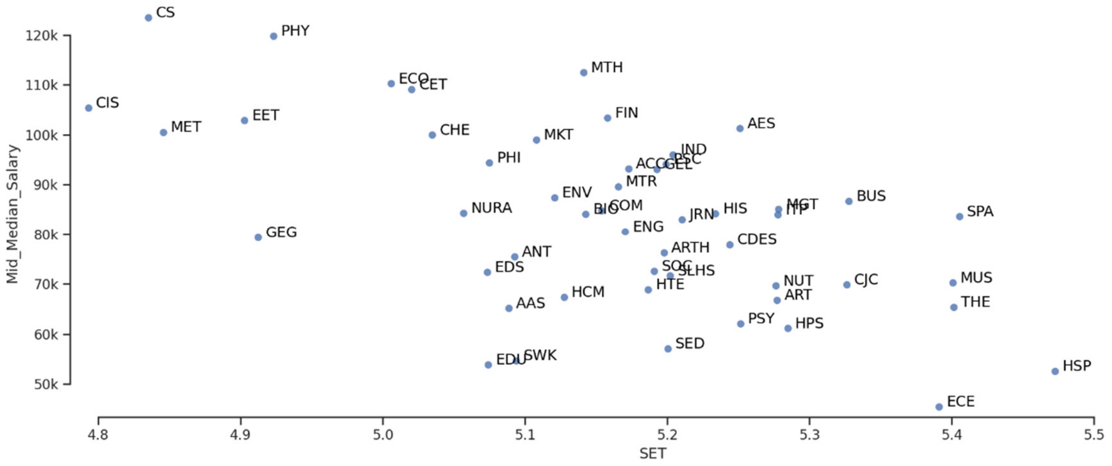


Figure 2 - Regression Line and Scatterplots for Sets an Mid-Career Income

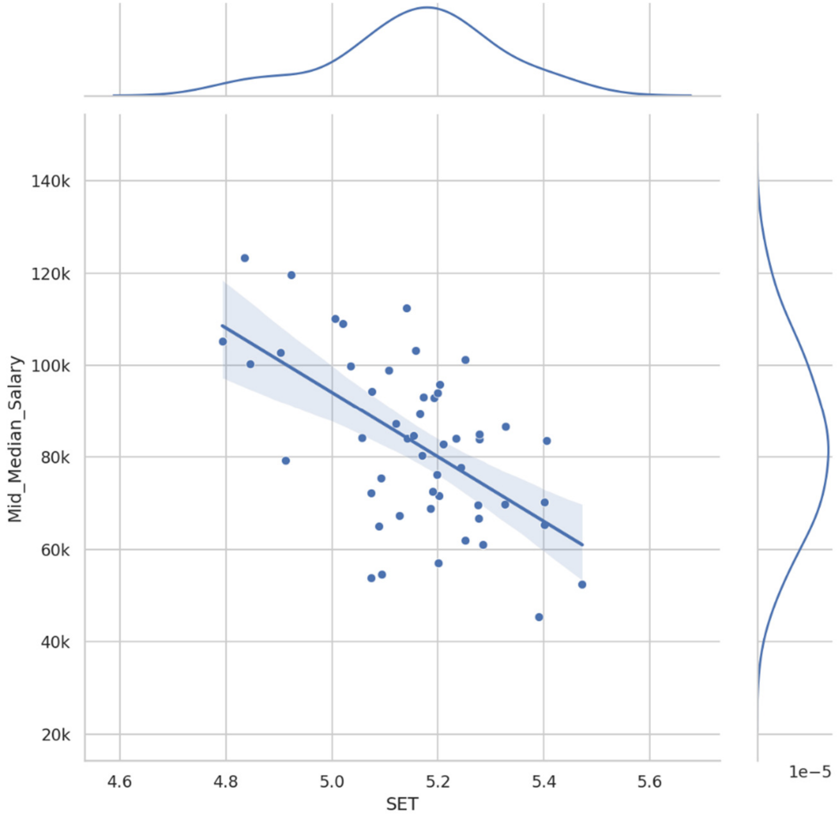


Figure 3 - Causal Diagram

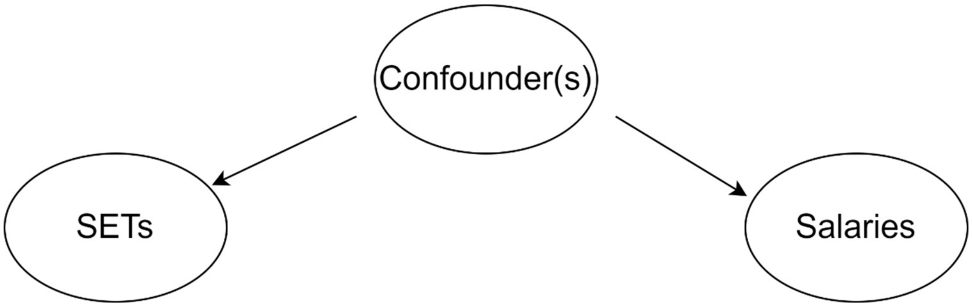


Figure 4 - Causal Diagram Including Quantitative Majors

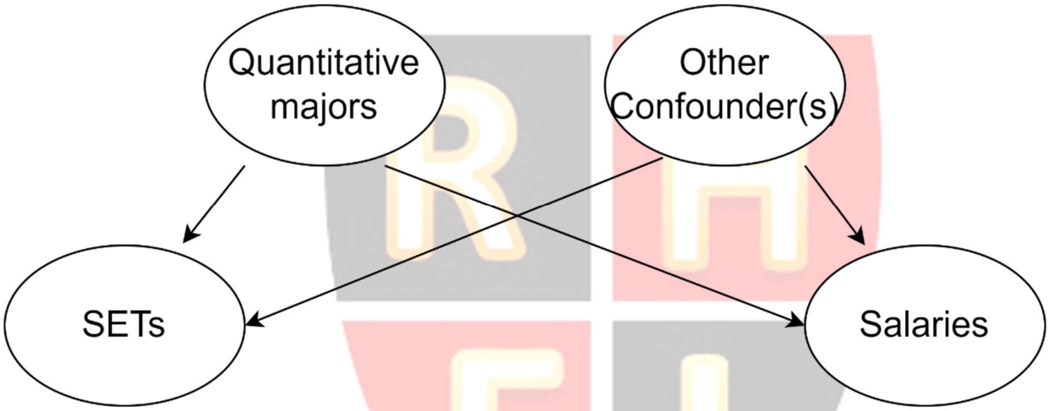


Figure 5 – Regression Line with Scatterplots for Set Mid-Career Income Including A Quantitative Indicator

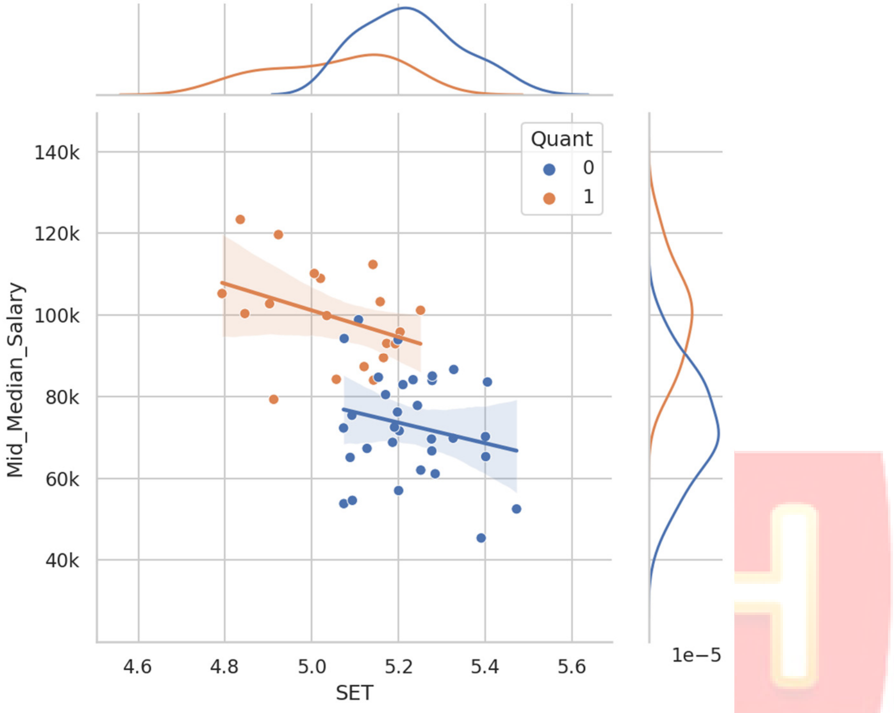


Figure 6 - Distribution and Means Of Set For Quantitative and Non-Quanitative Prefixes

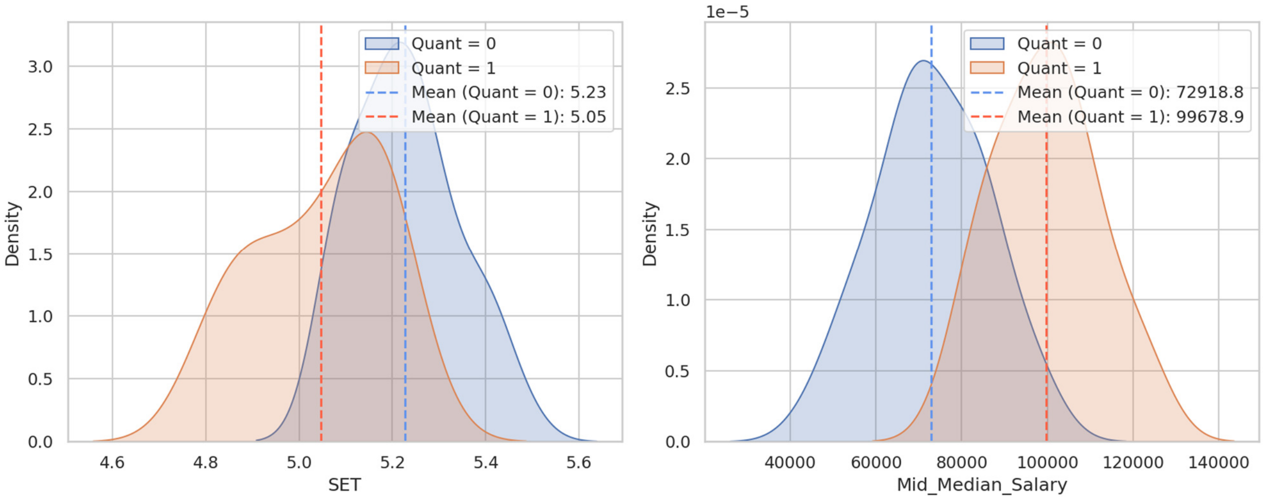
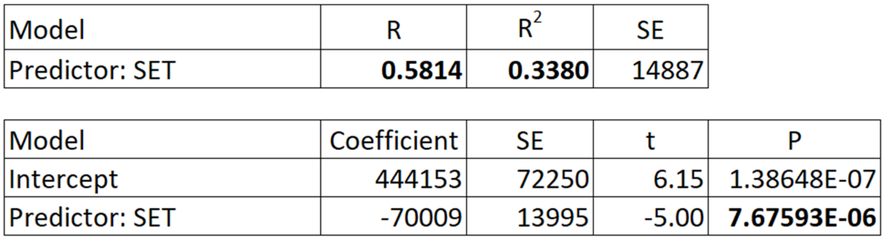


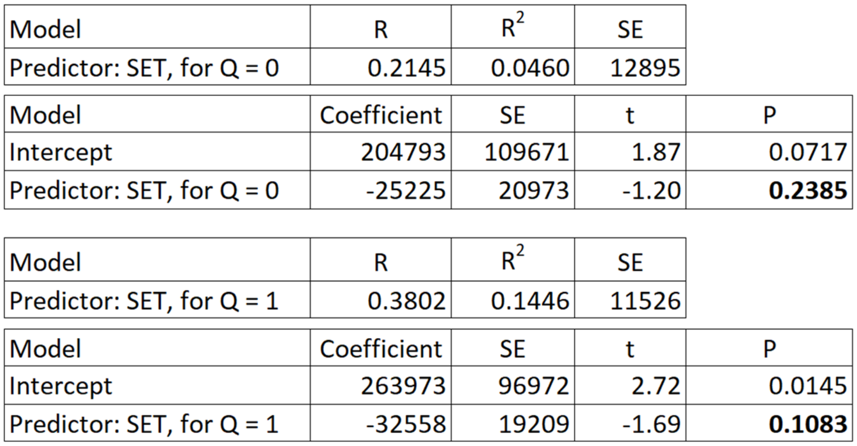
Table 1 - Model Summary and Coefficients for Set Vs. Salary



Model	R	R^2	SE
Predictor: SET	**0.5814**	**0.3380**	14887

Model	Coefficient	SE	t	P
Intercept	444153	72250	6.15	1.38648E-07
Predictor: SET	-70009	13995	-5.00	**7.67593E-06**

Table 2 - Model Summary and Coefficients for Set Vs. Salary Separated by the Quantitative Indicator



Model	R	R^2	SE
Predictor: SET, for Q = 0	0.2145	0.0460	12895

Model	Coefficient	SE	t	P
Intercept	204793	109671	1.87	0.0717
Predictor: SET, for Q = 0	-25225	20973	-1.20	**0.2385**

Model	R	R^2	SE
Predictor: SET, for Q = 1	0.3802	0.1446	11526

Model	Coefficient	SE	t	P
Intercept	263973	96972	2.72	0.0145
Predictor: SET, for Q = 1	-32558	19209	-1.69	**0.1083**

Table 3 - List of Prefixes in the Study

	Course Prep	Major	Quant
1	AAS	Africana Studies	0
2	ACC	Accounting	1
3	AES	Aviation and Aerospace Science	1
4	ANT	Sociology and Anthropology	0
5	ART	Art	0
6	ARTH	Art History, Theory and Criticism	0
7	BIO	Biology	1
8	BUS	Business	0
9	CDES	Communication Design	0
10	CET	Engineering	1
11	CHE	Chemistry and Biochemistry	1
12	CIS	Computer Information Systems	1
13	CJC	Criminal Justice and Criminology	0
14	COM	Communication	0
15	CS	Computer Science	1
16	ECE	Early Childhood Education	0
17	ECO	Economics	1
18	EDS	Secondary Education	0

19	EDU	Elementary Education	0
20	EET	Electrical Engineering Technology	1
21	ENG	English	0
22	ENV	Environmental Science	1
23	FIN	Finance	1
24	GEG	Geography	1
25	GEL	Geology	1
26	HCM	Health Care Management	0
27	HIS	History	0
28	HPS	Human Performance and Sport Theory	0
29	HSP	Human Services and Counseling	0
30	HTE	Hospitality, Tourism, and Events	0
31	IND	Industrial Design	1
32	ITP	Integrative Healthcare	0
33	JRN	Journalism	0
34	MET	Mechanical Engineering Technology	1
35	MGT	Management	0
36	MKT	Marketing	0
37	MTH	Mathematics and Statistics	1
38	MTR	Meteorology	1
39	MUS	Music	0
40	NURA	Nursing, Accelerated	1
41	NUT	Nutrition	0
42	PHI	Philosophy	0
43	PHY	Physics	1
44	PSC	Political Science	0
45	PSY	Psychological Sciences	0
46	SED	Special Education Courses	0
47	SLHS	Speech, Language, and Hearing Sciences	0
48	SOC	Sociology	0
49	SPA	Spanish	0
50	SWK	Social Work	0
51	THE	Theatre	0

Table 4 - Results for the Normality Assumptions and Welch`S T-Tests

Normality of Variable	**Shapiro-Wilk Test Statistic**	**P-Value**
Mid_Median_Salary, Quant = 0	0.9845	0.9154
Mid_Median_Salary, Quant = 1	0.9795	0.9366
SET, Quant = 0	0.9491	0.1356
SET, Quant = 1	0.9280	0.1593

Welch`s t-test / Variable	**Difference**	**t-value**	**P-value**
Mid_Median_Salary	0.18	4.79	3.82E-05
SET	-26760.2	-7.42	4.75E-09