Investigation Report

Question: Is there a relationship between the amount of money Year 12 Students earn on a weekly basis and the amount of money they spend at the canteen?

Introduction

Majority of Year 12 students have jobs or a source of income that they are able to spend at their own discretion. Spending of said income includes at the school canteen for small snacks or lunch. This report will present whether there is a correlation between the amount of money earned, and the amount of money spent at the canteen within the population of Year 12 students.

Considerations

The **explanatory variable** will be the amount of money the student earns per week.

The **response variable** will be the amount of money they spend at the canteen.

We will be using **a scatter plot** to display our data, and to observe whether there is a correlation.

Figuring the Sample Size

We decided that 50 students were a representative enough sample size of the general Year 12 population, rather than surveying the entire Year 12 cohort. Surveying the entire Year 12 population would be tedious and unnecessary as only using 50 unbiased and randomly selected students would be representative enough of the population – composing of 20% of the 250 students in Year 12. In order to remove bias and skewed data, the individuals were all randomly selected (random sampling), from both female and male genders. By ensuring that the respondents were all randomly selected it ensures a reliable representation of the Year 12 population.

Our Survey

The survey we designed consisted of 2 questions (Figure 1). The first question required to enter the amount of money earned each week either from allowances or a job. If they received no payments, they entered '$0'. The second question was also a short answer, requiring the student to enter the amount of money spent at the canteen, if applicable. We did not use any multiple-choice questions as it can skew our data due to its selective characteristics, limiting our representation of data.

By limiting the number of questions asked, it assisted on creating unbiased survey, allowing us to capture a diverse range of data, increasing the reliability and representation of our data to the general population of Year 12 students. Furthermore, by including students who don't have a job but receives allowances from parents, removes the biasness towards people with jobs only – thus creating a stronger representation of the population.

Having a clear and concise survey without unnecessary, biased questions provides reliable data we need to easily interpret, analyse, and develop a relationship between the amount of money earned weekly, and the money spent at the canteen.

Survey:
https://docs.google.com/forms/d/16Jq_2ERWPDheOmSh_znP5zOgUMbBpVUXE9mgqQxFjnU/edit

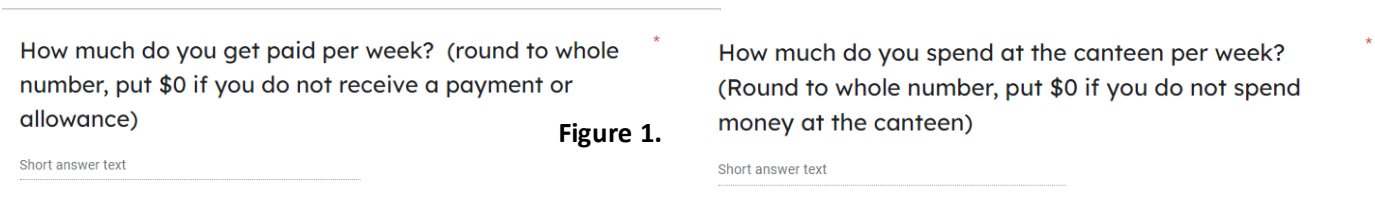


Figure 1.

Recording Data

Our survey was directly linked to a spreadsheet that contained all the responses. This process was entirely automatic, to eliminate any possible human error regarding transferring and recording the data.

Our data was placed into a table, with the explanatory variable (money earned) at the top representing the x axis, whilst the response variable (money spent) was placed at the bottom, representing the y axis, as seen in **Figure 2.**

Figure 2.

	1	2	3	4	5	6	7	8	9	10
Amount of Money Earned (x)	$340.00	$150.00	$560.00	$68.00	$210.00	$85.00	$67.00	$80.00	$150.00	$100.00
Amount of Money Spent at Canteen (y)	$25.00	$13.00	$10.00	$3.00	$16.00	$100.00	$3.00	$15.00	$8.00	$5.00

	11	12	13	14	15	16	17	18	19	20
Amount of Money Earned (x)	$85.00	$224.00	$80.00	$120.00	$145.00	$201.00	$129.00	$0.00	$170.00	$0.00
Amount of Money Spent at Canteen (y)	$3.00	$3.00	$10.00	$0.00	$0.00	$0.00	$0.00	$0.00	$3.00	$0.00

	21	22	23	24	25	26	27	28	29	30
Amount of Money Earned (x)	$0.00	$120.00	$20.00	$280.00	$109.00	$68.00	$120.00	$96.00	$112.00	$92.00
Amount of Money Spent at Canteen (y)	$15.00	$5.00	$0.00	$0.00	$3.00	$0.00	$15.00	$7.00	$3.00	$20.00

	31	32	33	34	35	36	37	38	39	40
Amount of Money Earned (x)	$0.00	$114.00	$116.00	$119.00	$100.00	$119.00	$105.00	$95.00	$111.00	$85.00
Amount of Money Spent at Canteen (y)	$4.00	$17.00	$5.00	$5.00	$4.00	$15.00	$5.00	$5.00	$0.00	$3.00

	41	42	43	44	45	46	47	48	49	50
Amount of Money Earned (x)	$165.00	$120.00	$100.00	$0.00	$98.00	$112.00	$90.00	$86.00	$80.00	$113.00
Amount of Money Spent at Canteen (y)	$20.00	$15.00	$0.00	$0.00	$4.00	$1.00	$10.00	$14.00	$3.00	$5.00

Spreadsheet:
https://docs.google.com/spreadsheets/d/1gQhB8kZpKo8Jc6Jue7hM7SZbzG5wmfTWRnzU3l-seNI/edit?usp=sharing

Procedure

1. Due to the year group having more than 100 students, with only a small portion of the population actually earn an income and spend money at the canteen on a regular basis, a group of 30 students were selected to participate in the survey.

2. The developed survey consisting of two questions were sent randomly to a group of students. Students were randomly selected by going to different areas of the school in order to prevent a bias towards a proportion of the population.

3. All responses to the survey were immediately tabulated on a spreadsheet that would later assist us on statistical calculations, such as finding the averages and minimums/maximums.

4. Knowing that the money earned is the explanatory variable, and the amount of money spent is the response variable, we are able to create a scatter graph in order to visually represent the data we have collected, allowing us to see if there is a correlation between the explanatory and response variable.

5. Via the assistance of our CASIO Classpad, we are also able to calculate the correlation coefficient and the determination of correlation easily by simply entering the data onto the Classpad spreadsheet.

Results and Discussion

Our data can be seen in **Figure 2**, with all 50 responses tabulated onto a spreadsheet. **Figure 3** illustrates the data plotted onto a scatter graph, showing **no correlation**, alongside with some outliers.

Figure 3.

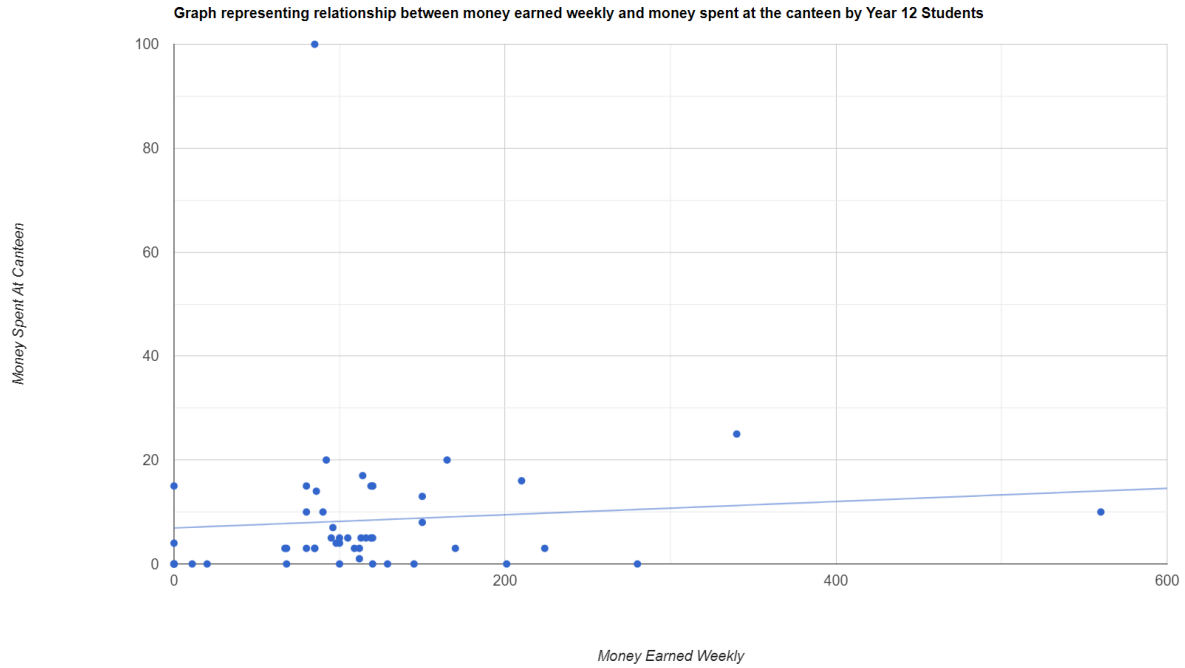
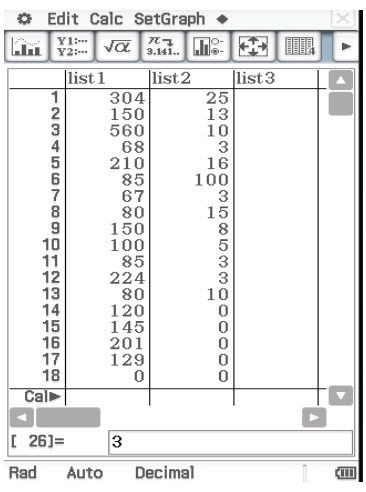


Figure 3 shows the relationship between the Money earned by Year 12s on a weekly basis, and the money they spent at the canteen. On the graph, the money earned weekly is represented on the x axis as it is the explanatory variable and the money spent at the canteen is represented on the y axis, as it is the response variable. All 50 responses were plotted, including outliers. The line of best fit (regression line) is also constructed in the graph, which shown as a weak positive correlation, however the data points themselves show no correlation. Furthermore, the negligible correlation of the data points is also affected due to the outliers of points: (100,85), (280, 0), (340, 25) and (560, 10) as they observed to have an abnormal distance from the rest of the population. Having outliers in an investigation is almost unavoidable, as not all students are the same – where some may get paid more than others whilst some may spend less at the school canteen. However, in order to ensure unbiasedness and enforce the reliability of our investigation, we included the outliers within our graph. The average earned by a Year 12 is $118 per week, with an average of $8 spent at the canteen.

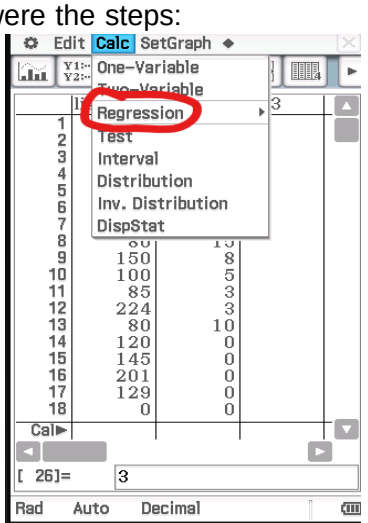
Due to there being no correlation determined, we **cannot** assume the pattern that the more the students earn, the more they spend at the canteen is true nor develop a correlation between the explanatory and response variable.

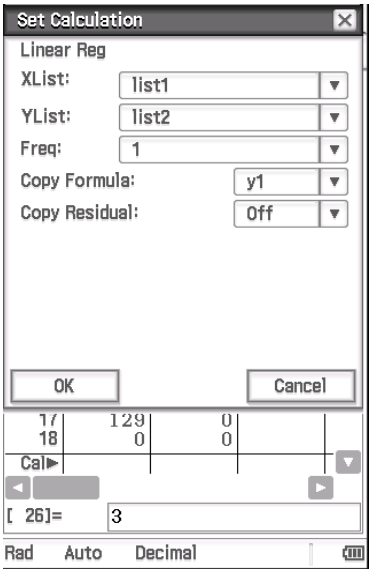
The utilisation of the CASIO Classpad was implemented to assist us to calculate the correlation coefficient (r) and the coefficient of determination (r^2), alongside to find the linear equation of the regression line in the scatterplot. In order to find the correlation coefficient using the Classpad, these were the steps:



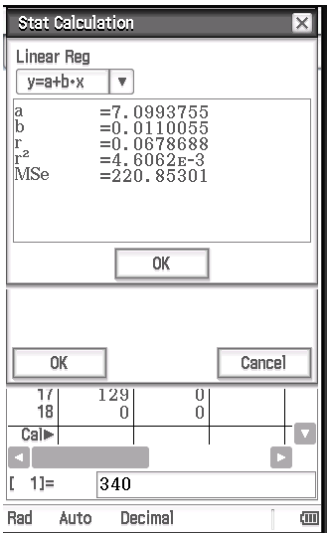
1. Place all data in the Classpad spread sheet. X values in List 1, Y values in List 2.

2. Click on ‘Calc’ and press regression. A subsection should appear and click ‘Linear Reg’





4. Keep all functions the same, except change 'Copy Formula' to y1. Press 'ok'



3. A window should appear, showing you a value, b value, correlation coefficient $(\boldsymbol{r})$ and the determination of coefficient $(\boldsymbol{r^2})$.

Depending on what formula is selected, a and b may represent different things.

In this instance, a represents the y intercept and b represents the gradient.

As a result, we can decide that the correlation coefficient is 0.0678688 and the coefficient of determination is 4.6062e-3 (0.0046062).

Our linear regression equation $(\hat{y} = a + bx)$ is: **ŷ = 7.0993755 + 0.0110055x.** Through this equation's gradient *(0.0110055x),* we can observe that for every $1 earned, $0.0110055 is spent. When no money is earned $(\boldsymbol{x = 0})$, $7.10 will be spent (rounded to the nearest 5 cents).

Correlation Coefficients $(\boldsymbol{r})$

Knowing the correlation coefficient is always between -1 and 1, we can determine that:

- -1 means there is a strong negative relationship.
- 1 means there is a strong positive relationship.
- 0 means there is no correlation.
- Approximately between 0.6 and 0.7 there is a moderate correlation.

In order to justify the no correlation relationship between the explanatory and response variable, we compare our correlation coefficient of 0.0678688 to the threshold brackets mentioned above. Though our coefficient is lower than 1, meaning it is not a strong positive correlation, it is still above 0 – however not majorly above 0. Due to this we will just assume there is **no correlation** as the correlation coefficient of 0.0678688 is visually unobservable to the eye – however *theoretically*, there is a very weak positive correlation.

To further justify the no correlation of our data, according to ***Pearson's correlation coefficient*** table, the correlation coefficient of 0.0678688 falls under the category of no correlation (values between 0.00 and 0.30 or 0.00 and -0.30)

Size of Correlation	Interpretation
.90 to 1.00 (−.90 to −1.00)	Very high positive (negative) correlation
.70 to .90 (−.70 to −.90)	High positive (negative) correlation
.50 to .70 (−.50 to −.70)	Moderate positive (negative) correlation
.30 to .50 (−.30 to −.50)	Low positive (negative) correlation
.00 to .30 (.00 to −.30)	negligible correlation

Coefficient of Determination (r^2)

Knowing that the determination of coefficient always falls between 0 and 1, we know that the higher the determination of coefficient is, the better the regression line suits the data population, thus results in a strong relationship.

As our coefficient of determination (0.0046062) is above 0, it shows that the regression line weakly supports the data population. Due to this**, we can conclude that 0.46% of the variation in the money spent at the canteen can be explained by the variation in the money earned weekly by Year 12 students.** By having only 0.46% of the data fit the regression line, it reinforces the idea that the regression line weakly represents the data population, resulting in **an unreliable** model for future predictions.

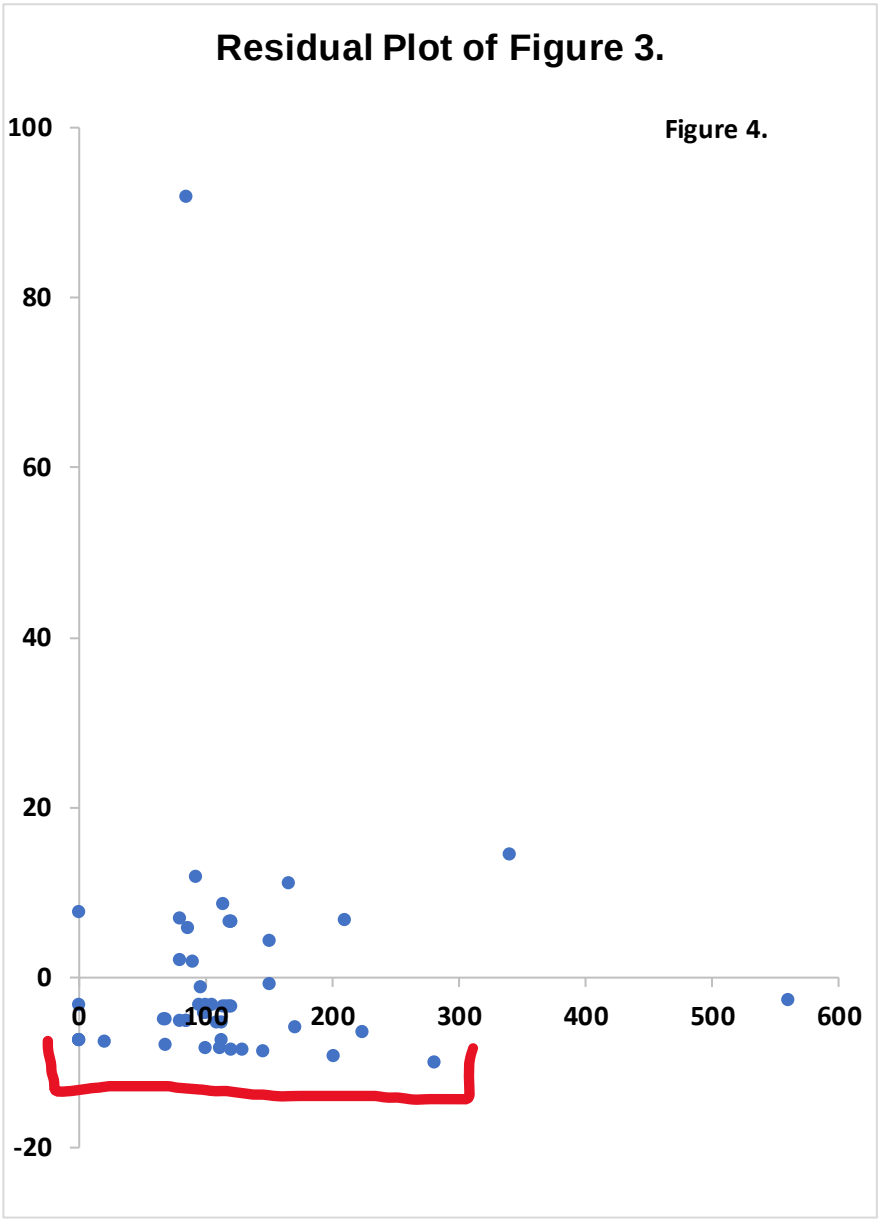


Figure 4.

Residual Plot

Using the line regression equation we obtained earlier, we can construct a residual plot.

As seen in **Figure 4**, it is evident that the data **does not fit a linear model**, due to the lack of correlation in the data points.

Majority of the data points are also condensed to the left, further emphasising the non-linear characteristic of the data.

Making Predictions

As we have obtained our linear regression equation from the classpad, we can now successfully make predictions based on the equation.

There are two types of predictions:

- **Interpolation:** predicting values that fall within our data range, more reliable as we can reference off to similar data points that can be found within our data range.
- **Extrapolation:** predicting values outside our data range, less reliable as we cannot accurately describe nor explain data that we cannot observe.

Our data range for the money earned each week is between **0 and 560**. Any predictions for values between these two are considered to be interpolated. Any values outside these values are considered extrapolated, thus unreliable.

Interpolated Prediction

Equation: $\hat{y} = 7.0993755 + 0.0110055x.$

where x = \$120

$\hat{y} = 7.0993755 + 0.0110055(120)$

$\hat{y} = 8.4200355$

∴ We can reliably predict that if a student was to earn \$120 per week, they would spend \$8.45 at the canteen (rounded to the nearest 5 cents).

Extrapolated Prediction

Equation: $\hat{y} = 7.0993755 + 0.0110055x.$

where x = \$670
$\hat{y} = 7.0993755 + 0.0110055x.$ (670)

$\hat{y} = 14.4730605$

∴ We can unreliably predict that if a student was to earn \$670 per week, they would spend \$14.50 at the canteen (rounded to the nearest 5 cents).

Association and Causation

As there is no association between the money earned and the money spent at the canteen, we can determine that the more money earned on a weekly basis does **NOT** cause an increase of spending at the canteen.

However, if there **WAS** a correlation, the amount of money earned would not entirely be responsible for the causation of the amount of money spent at the canteen – as it can be influenced by other variables that are out of our control. Thus, we can conclude that **association ≠ causation.**

These variables include:

- Forgetting lunch, resulting in increased spending at the canteen
- Expensive pricing, drawing away people to spend money at the canteen
- Not bringing enough food/drinks to suffice hunger needs, increasing spending at the canteen.
- 'Stingy' students, wanting to save money, decreasing spending at canteen.

Conclusion

According to our recorded data of a sample size of 50 students representing 20% of the 250-student cohort, we can easily observe that there is **no correlation** between the amount of money a student earns weekly to the amount of money spent at the canteen.

Aforementioned, there are multiple other variables that affects the amount of money that student will spend at the canteen, not only just how much they earn each week. Thus, we can also state that the amount of money earned by the student on a weekly basis is **NOT** a causation to the amount of money spent at the canteen.

However, keep in mind that this investigative process was never entirely perfect – there will always be some sort of bias or error within our data population. Bias may include students over-stating their weekly wage/allowance in order to not feel embarrassed compared to other student's weekly wages and under-stating the true amount of money spent at the canteen to avoid embarrassment, showing the money 'wasted' at the canteen. For example, an **extreme responding bias** (meaning they provide an extreme or 'exaggerated' response) is evident where a student state they earn $560 in a week, which is a considerable amount over the average wage of a 16–17y.o student. Errors within our investigative processes includes having to alter and amend the survey question when we already had responses coming in. This as a result contributes to the bias, as not all student respondents had answered the same amount and type of questions, possibly skewing our data, making it unreliable. Our sample size however was entirely randomly selected, each representative for different populations of the Year 12 cohort, which as a result increases the validity of the investigation.

There are various improvements we can incorporate in future investigations, the main one being ensuring we have finalised our questions prior release to respondents. This as a result makes the survey clear and fair for everyone and prevents creating a bias towards certain students as some may answer different questions to others. Some other improvements we can incorporate in regards to reliability includes increasing our sample size, ensuring that it is more representative of the Year 12 cohort, rather than only 20%. Increasing our sample size also reduces the effect of bias and errors made – however does not entirely remove the impact. In order to prevent the overstatement and understatement of money earned and money spent, we can also request for physical proof of the expenses made by the student, in order to reinforce their statement of how much they earn and spend – though it may seem overdoing but is necessary if we wish to eliminate biasness within our data.