My exploratory journey started with questions correlation using a regression line to determine the r-squared that would tell me the strength of the relationship of two quantitative variables thus deciding if the model could be used for prediction.

Initially, I wanted to answer my questions with three columns: Age, Purchase Amount (USD), and Review Rating from the original DataFrame. First, I was curious to find the average age in the dataset (44 years old). Then, I calculated how many purchases were made by age and how many review ratings were made by age. But I needed averages to plot the regression lines, not the counts.

I had to opt for data engineering by adding two new columns to the original DataFrame to plot the regression lines. The new columns were Purchase Amount (USD)\_avg by Age and Review Rating\_avg by Age; then, I merged the two new means into the DataFrame.

With a similar hypothesis, I plotted the two regression lines, and both had an r-squared equal to zero and correlation coefficients of zero, as shown in Figures 1 and 2.

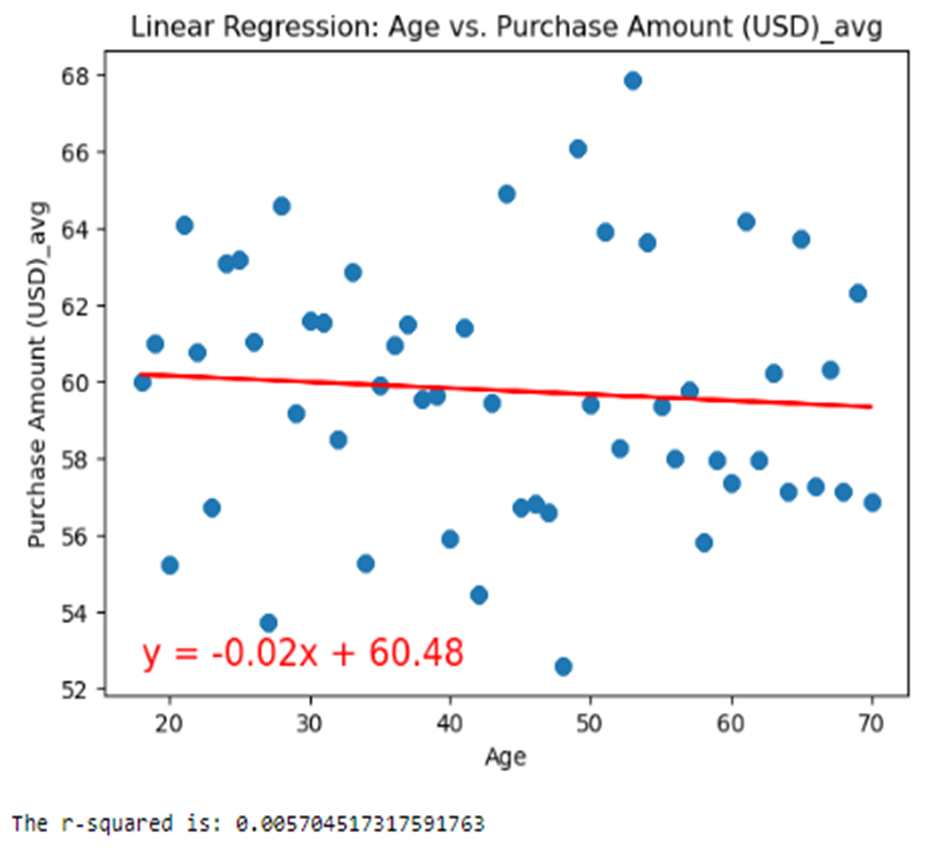


Figure 1

Hypothesis: Age impacts Purchase Amount (USD).

The scatter plots indicate no relationship between the variables, no relationship between Purchase Amount (USD)\_avg and age.

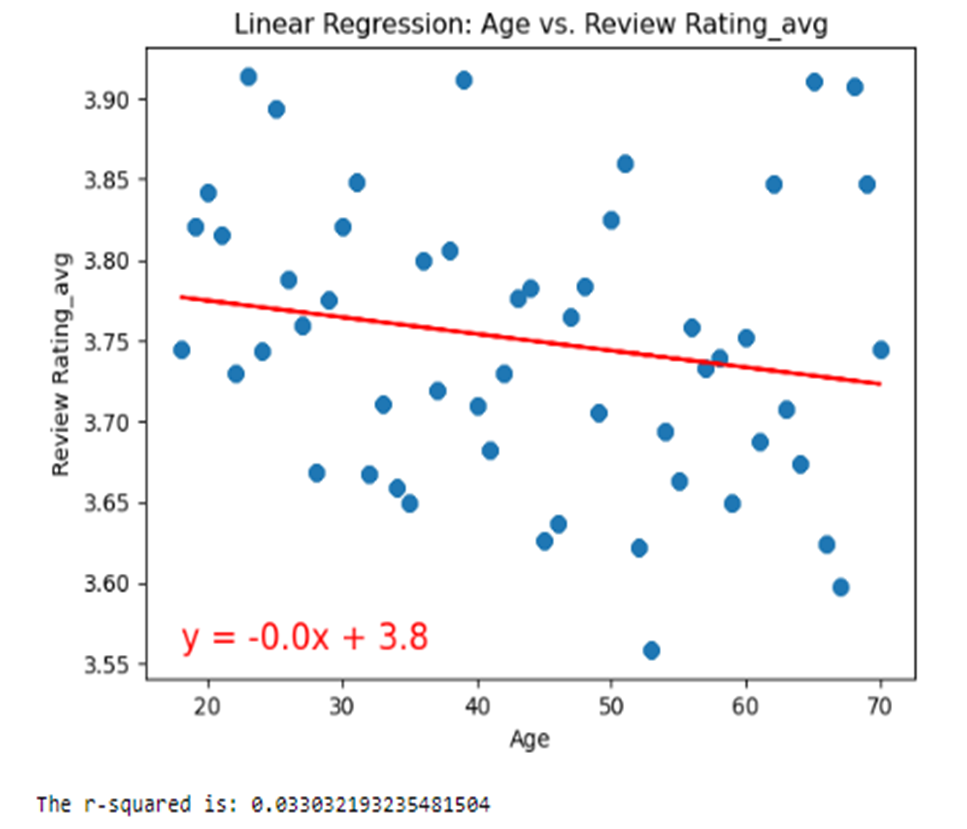


Figure 2

Hypothesis: Age impacts Review Rating.

The scatter plots indicate no relationship between the variables, no relationship between Review Rating\_avg and Age.

We could not use either line model for prediction. I calculated the mean and median for both datasets to verify my results. For Age vs. Purchase Amount (USD)\_avg, the mean = 59.76 and the median =59.66, while the mean and the median for Age vs. Review Rating\_avg were mean =3.748 and the median = 3.744. In both cases, the mean and the median were almost the same. Both graphs have a normal distribution. In a normal distribution, the means lie on the line, the points centered on the line or closer to the line. These results proved that the regression lines were correct at establishing that the variables had no impact on each other. Furthermore, the points on the scatter plots do not show any pattern; there is no relationship between them.

Later, I used a Ttest to compare the two means of the two new columns. However, I could not compare the two averages to determine a difference between the variables because the mean and median of both averages were very close as discussed above (the Average Purchase Amount by Age mean = 59.76, median = 59.66, and Average Review Rating by age mean = 3.748, median = 3.744).

I had to find two groups (not averages, not both quantitative) to determine whether the two groups' means were different. I decided on the group of males and the group of females. Gender against Purchase Amount (USD) and Gender vs Review Rating. See figures 3 and 4 below.

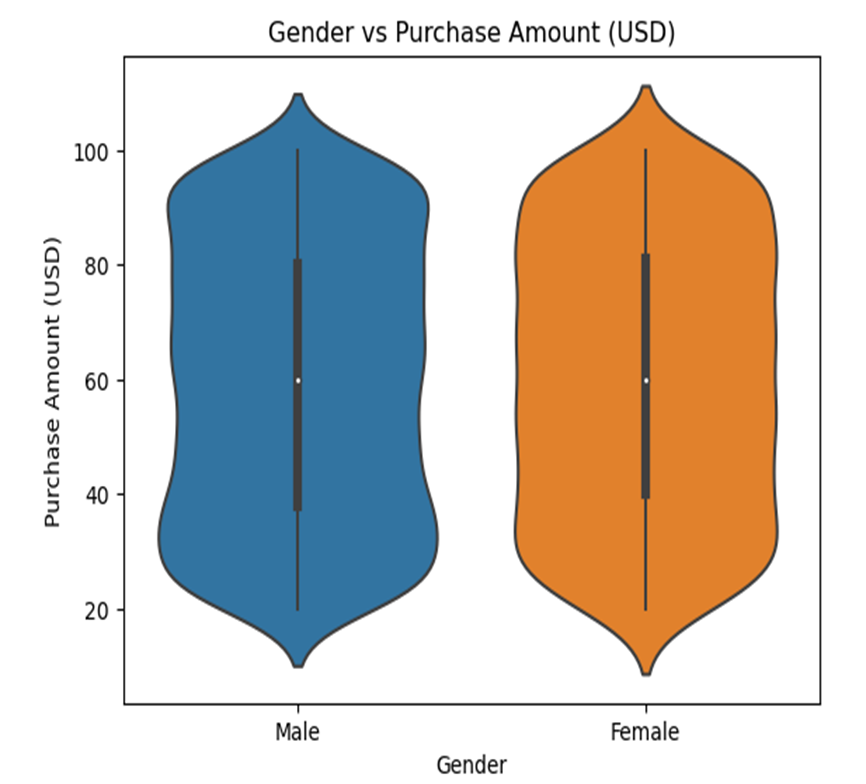


Figure 3

My null hypothesis was Gender means in Purchase Amount (USD) are equal. The alternative hypothesis was Gender means in Purchase Amount (USD) are not equal.

Using the t-test to compare the two means resulted in Gender and Purchase Amount with a p\_value = 0.38.



Figure 4 (p\_value = 0.61

My null hypothesis was Gender means in Review Rating are equal. The alternative hypothesis was Gender means in Review Rating are not equal.

Using the t-test to compare the two means resulted in Gender vs Review Rating p\_value = 0.61.

When compared to a significance level of 0.05, both p\_values are higher, which indicates results are not statistically significant for both datasets. I calculated the mean for both groups to verify my results. The two group means for either the male group mean (59.53) or the female group mean (60.24) in the Purchase Amount (USD) are very close. The means for the male group (3.75) and female group (3.74) in Review Rating are also very close.

**Findings**

The findings are supported by correlation with a regression line, mean and median numbers, and tTest results.

There is no correlation, no predictive power between Purchase Amount (USD)\_avg and age neither by Review Rating\_avg and Age as shown by the two almost flat regression lines, there is no ability to predict whatsoever. Both models do not fit the data well.

I failed to reject both null hypothesis Gender means in Purchase Amount are equal and Gender means in Review Rating are equal. Both male and female group means in Purchase Amount, and male and female group means in Review Rating are the same. The difference between the two groups is not [statistically significant](https://www.scribbr.com/statistics/statistical-significance/). We assume that it is very likely that there is no difference in male and female groups in both Purchase Amount (USD) and Review Rating.

**Conclusions**

The regression lines are not a good fit to predict the relationship between age, the independent variable, and either Purchase Amount (USD)\_avg or Review Rating\_avg (dependent variables).

The tTest resulted in a p\_value greater than 0.05, which is not [statistically significant](https://www.scribbr.com/statistics/statistical-significance/) and indicates strong evidence for the null hypothesis. In other words, I failed to reject the null hypothesis, which means I retained the null hypothesis and rejected the alternative hypothesis. The sample data does not provide sufficient data to conclude that there is a difference between the two groups of means.

**what are the big ideas? actionable insights?**

There is no correlation between Age and either Purchase Amount (USD)\_avg or Review Rating\_avg

The male and female group means in Amount Purchase (USD) or Review Rating are likely equal.

**3 bullet points that we should be taking away from this**?

Age and Gender are independent variables with no significant effect on other variables.

There is no need to establish marketing strategies based on customer age or gender preferences.

**The limitation** of the study was the size of the data. I would have preferred a larger size of data; more data would have been more certain data to refute the null hypothesis.

**what would you do differently**?

A larger sample size would have probably identified any outliers in the dataset and would be more likely to obtain statistically significant results to generalize to the population.

we would not recommend retailers base their market strategy on consumer demographics such as age or gender

**Data:**

Average Purchase Amount (USD) by Age mean = 59.76

Average Purchase Amount (USD) by Age median = 59.66

**Data:**

Average Review Rating by age mean = 3.748

Average Review Rating by age median = 3.744

A diagram of fat intake and fat intake

Description automatically generated

When there is no relationship between x and y, the values of x are of no help in predicting the y scores, so we might as well use the mean of y, or to predict y scores. In the left-hand side of the above figure, there is a flat line drawn at the mean. The best way to predict the y scores is with the mean of y. To the extent that there is a relationship between x and y, there will be some slope in the line of prediction. So, the degree to which the regression line is sloped compared to the mean, represents the amount we can predict y scores.

The assumption of normality is on the residuals, not the independent or dependent variables.

A diagram of a fat intake

Description automatically generated

 by the little vertical lines I've drawn in the above figure