To answer the questions proposed by Turtle Games, I began by making predictions using Regression. After importing and printing the data to understand the contents and layout, I stored it in a dataframe ready for analysis. I removed the redundant columns 'language' and 'platform'. I also simplified the column names for remuneration and spending score to make it easier to reference later in the code.

Next was defining the variables for the linear regression where the dependent variable (y) was 'loyalty points' and the independent variable (x) was 'age'. These variables had a no comprehendible correlation. The OLS test also did not output a strong correlation between the two variables.

I then looked at the relation between loyalty points and remuneration. The linear regression shows a strong positive correlation - the higher the remuneration for a customer, the higher the loyalty points. This can be explained due to higher remuneration resulting in higher disposable income for a customer for a luxury good such as games. The OLS test confirmed this with a positive trend line.

Lastly, I looked at the correlation between loyalty points and spending score. The regression model displayed a positive correlation with the higher the spending score, the more loyalty points customers had. The OLS test further confirmed this with a positive regression line. We can use predictive regression model to suggest that a customer with higher remuneration will collect more loyalty points and the more they spend the higher their spending score and therefor the higher the loyalty points. Customers can collect loyalty points by spending more at Turtle Games.

I then utilised Clustering to make prediction. Upon importing the necessary libraries, I created a new dataframe with just remuneration and spending score. I then plotted a scatterplot and pair plot for the dataframe. To understand the optimal number of clusters, I used the elbow method which showed that 5 clusters will be ideal. The silhouette method had a peak at 5 clusters, also showing this number would be an optimal number of clusters. I created multiple pairplots and scatterplots with 6, 5 and 4 clusters to confirm 5 was the optimal number. I was then able to identify 5 distinct groups within the data between remuneration and spending score – I.e. there were 5 main groups with different spending habits/remuneration/spending scores which can all be marketed and can be used to target specific market segments.

My final analysis in python was that of customer sentiment with reviews. I focussed on the review and summary columns of the dataframe and looked at the descriptive statistics of the two columns. From the describe () function, I could see that the words 'love' and 'it' were the top words in the review column with 'five' and 'stars' were the top words in the summary column. Although this is not as in-depth as I would've liked, this provides some insight into the marketing campaigns. The top words from each column would be effective if used in the marketing campaigns as you are speaking directly to Turtle Games demographic and speaking the customer's language. To improve my sentiment analysis, I would look to go further with my code here.

I then started work on the sales data in R. I imported the data and dropped the 'ranking', 'year', 'publisher' and 'genre' columns as they were redundant. I used the histogram, boxplot and scatterplots in the qplot function to see which would display the data best between product, platform and sales. As the products are named by number, they were plotted in a numerical range on the x axis in my plots. The graph that displayed the most telling data was the scatter plot for product vs Global sales. This showed a negative hyperbola shape. The lower the product number the more sales it predicted.

I then began to investigate how reliable the data was. I first created a Q-Q plot for global sales and added a line. I then went onto perform the Shapiro-wilk test which showed a p-value < 2.2e-16 – as this is very close to 0, this means the null-hypothesis is rejected and my test is statistically significant. The skewness was 3.042357 meaning the data is highly skewed with a heavy tail meaning the data is not evenly distributed. This could be due to how the data is collected or the sample size of the data. This is something I would like to investigate further. By checking for kurtosis, I got the value 17.6566 which means the data has a heavy tail and could potentially highlight lots of potential outliers. I then plotted the skewness and kurtosis of the data which again, showed a heavy-tailed distribution. Although the data is statistically significant and the null hypothesis has been rejected, I believe by gathering a larger sample size for the data collection and will ensure the data is more symmetrical in its normal distribution. This will ensure there are no outliers, and the data is more reliable.

Finally, my last analysis in R was to create a linear regression model. The p-value was 2.2e-16 Meaning the null hypothesis was rejected. The R^2: 0.3722 value means showing. Ahigh level of correlation between the sales in Europe, America and Globally.

My recommendations to Turtle Games would be to target the 5 cluster I have identified through their marketing campaigns, with even focus across Europe and the US. Their goal should be to get customers to spend more as well as increasing the customer base so they can collect more data from sales to ensure the data is not heavily tailed. This will also mean the customers will gain more loyalty points which can generate repeat business. The products that are labelled by the lower numbers generate the most sales – if this is the price of the products this means if they wanted to have more evenly distributed data, they should focus on pushing the higher priced products. Words such as 'love it' and 'five star' should be used in the marketing campaigns as they will be speaking the customer's language and will generate more custom according to the sentiment analysis. Improvements to my data analysis can be made through more in-depth sentiment analysis as well as a more evenly distributed data set for sales generated by Turtle Games.