Assignment: Predicting Future Outcomes

Turtle Games seeks to improve its sales performance by leveraging customers trends and has contracted us, a team of data analysts, to draw insights from its sales data. The objective of our team is to analyze collected data to understand loyalty points accumulation, targeted market segments, social data utilization, product impact on sales, data reliability, and regional sales relationships.

The analysis was conducted in Python and R to import, clean, and analyze data. Libraries such as pandas, numpy, seaborn, and sklearn were used for data manipulation, visualization, and machine learning. Data preparation included removing unnecessary variables, dealing with missing values, and converting data types. Linear regression models, k-means clustering, NLP, and descriptive statistics were employed to answer Turtle Games' questions.

For loyalty points accumulation, an OLS Linear Regression model was used to analyse the relationship between the variables. The multiple linear regression model was also employed to predict loyalty points based on the independent variables. Moreover, it was determined that there was no multicollinearity between the variables signifying the stability of model. Spending and remuneration were identified as significant factors in loyalty points accumulation, while age was deemed less influential. The marketing team can tailor their campaigns taking into account people’s spending and remunerations.

To identify customer segments, k-means clustering was used on remuneration and spending data, with the elbow and silhouette methods determining the optimal number of clusters. Five distinct clusters were identified, providing insights for targeted marketing efforts such as for high-income, high-spending customers (Cluster 2) might be targeted with premium products or exclusive offers, whereas low-income, high-spending customers (Cluster 1) could benefit from budget-friendly deals or financing options. The distinct groups are shown in the table:

Cluster	Observations	Description	
		Remuneration	Spending Score
0	774	Mid	Mid
1	269	Low	High
2	356	High	High
3	271	Low	Low
4	330	High	Low

The k-means clustering approach has some limitations, however, such as sensitivity to the initial placement of cluster centroids and the assumption of equal-sized and spherical clusters, which may lead to suboptimal clustering results or difficulty in identifying the true underlying structure in the data.For further improvements to the model, incorporating other customer attributes, such as demographics or purchase history, could help refine the customer segmentation and provide a more comprehensive understanding of the customer base.

For social data analysis, NLP techniques were applied to customer reviews, extracting polarity and subjectivity scores. After removing stop words like “and” and “the”, the word clouds generated from customer reviews highlighted frequently mentioned words such as “great” and “fun”. Customer reviews and summaries were predominantly positive, with higher subjectivity scores suggesting a stronger influence of personal opinion. If a common word in negative reviews such as “difficult”, the marketing team could address the issue by finding the product in question and highlighting alternatives or a solution in their campaign. The limitations of the NLP approach,

however, is that there is a high potential for misinterpretation due to context, sarcasm, or the need for additional text preprocessing (e.g., stemming, lemmatization). This is evident in the model as there are certain reviews that are labelled negatively but clearly have a positive sentiment given the context.

The impact of products on sales, data reliability, and regional sales relationships were analysed using descriptive statistics, Q-Q plots, Shapiro-Wilk tests, and correlation matrices. A stand-out product in particular was product 107 which had the highest values for NA_Sales, EU_Sales and Global_Sales. These stand-out products should be analysed more to find any patterns behind their immense success such as whether the most successful games belong to a certain gaming platform like "Wii". Moreover, the top 10 products show that sales in the NA region are consistently higher than in the EU region. The Q-Q plots suggest the data is not normally distributed as a large number of points deviated from the reference line. The sales data was determined to be not normally distributed based on the skewness and kurtosis values. For further analysis, non-parametric tests as well as transforming the data can be considered to remedy the fact that the data is not normally distributed.

The correlation matrix showed that there are positive correlations between all pairs of sales data columns, with the strongest correlation between NA_Sales and Global_Sales and that generally when sales increases in one region, they are likely to increase in other regions as well. This point is further supported when observing the heatmap which shows a darker shade of blue for NA_Sales vs Global_Sales compared to EU_Sales vs Global_Sales. The relationships between regional and global sales were explored further using multiple linear regression models. The p-values for NA and EU were both below 0.05 indicating that they are statistically significant. Furthermore, both variables had very low standard errors which indicates the estimates of the regression coefficients are precise and reliable. The model yielded a very high R-squared value of 0.9668 which suggests the model has very strong explanatory power. The table below shows the accuracy of the model's prediction:

Actual Global Sales	Predicted Global Sales	Percentage Difference (%)
67.85	68.06	0.30
3.06	7.36	21.80
4.32	4.91	13.62
3.53	4.76	34.87
23.21	26.63	14.72

Overall, the multiple linear regression model seems to perform well in most cases, with only one relatively weak prediction. This suggests that the model is generally strong and provides useful predictions for Global_Sales. To further improve the model, we could consider additional variables that may influence global sales such as marketing efforts or product release dates.

In conclusion, the data analysis provides valuable insights into Turtle Games' customer trends and sales performance. By understanding the relationships between regional sales and global sales, the company should prioritize increasing sales in both North America and Europe to maximize overall performance. This can be achieved through tailored marketing efforts for each region, considering regional gaming preferences and cultural differences, and creating products that cater specifically to these markets.

Furthermore, identifying distinct customer segments through k-means clustering enables the company to target its marketing efforts more effectively, catering to the needs of high-income, high-spending customers as well as budget-conscious gamers. Additionally, the insights from NLP analysis of customer reviews can help the company pinpoint potential issues or areas of improvement for its products, directly addressing customer concerns and enhancing satisfaction.

Overall, the provided insights and predictions create a solid foundation for data-driven marketing strategies and sales improvement. By leveraging the knowledge gained from this analysis,

Turtle Games can better target specific market segments, capitalize on regional sales trends, and boost its overall performance. It is essential for the company to continuously refine and update these models, incorporating new data and variables, to ensure the ongoing effectiveness of its data-driven decision-making process.