# Section One – Amazon Sales and Review Data Analysis

## Research Question

How can Amazon sales and review data be analysed to extract meaningful business insights?  
  
This study aims to evaluate whether structured product metadata — including pricing, customer reviews, and ratings — can be used to model sales patterns and extract insights that inform business decisions in e-commerce contexts.

## Executive Summary

The objective of this project was to examine how structured Amazon UK product data can reveal insights about short-term sales performance. The dataset comprised over 2.2 million products, each described by attributes such as review volume, average star ratings, price, category, and estimated monthly purchases. These were analysed using Python in Jupyter Notebook, applying data cleaning, feature engineering, exploratory analysis, and machine learning regression models.  
  
The models developed included linear regression and Random Forest regression. While performance metrics such as Mean Absolute Error (MAE) showed that Random Forest improved upon the baseline, the R² score remained low (0.021), indicating limited explanatory power. Nonetheless, visual and statistical analysis identified review-related features and bestseller status as the most influential factors.  
  
The business value lies in identifying product-level traits that correlate with sales, particularly review engagement and visibility status. These insights can help sellers prioritise inventory, price competitively, and allocate promotional resources. Recommendations are offered for future improvement, including sentiment analysis and time-series modelling.

## Introduction and Project Background

Amazon has become a critical marketplace for sellers worldwide, with millions of active listings in categories ranging from electronics to household items. Understanding what drives consumer purchasing behaviour on such platforms is essential for competitive advantage. This project explores whether structured product metadata can meaningfully explain sales performance, using statistical and machine learning approaches.  
  
The broader context lies in the increasing role of data-driven retail, where sellers rely on automation, recommendation engines, and predictive models. By leveraging review scores, prices, and categorical information, businesses can optimise decisions such as pricing strategy, product bundling, and promotional campaigns.

## Importance of the Project

The ability to extract and interpret behavioural signals from structured data is of increasing importance in e-commerce. Companies like Amazon, eBay, and Etsy routinely track customer interactions to optimise listings. Previous studies have shown that customers are heavily influenced by peer reviews and visual indicators of quality and popularity (Chen et al., 2004). This project adds value by replicating such analysis at scale using publicly available data and accessible machine learning tools.  
  
From a practical perspective, this analysis can inform better marketing, more accurate forecasting, and smarter inventory management — essential functions for online retailers and third-party sellers.

## Literature Review

Brynjolfsson et al. (2003) coined the term “Long Tail” to describe how online retailers can profit from a wide range of niche products, rather than focusing solely on bestsellers. This has shaped Amazon’s strategy of offering an extensive inventory. Chen et al. (2004) found that review count and star rating both significantly affect consumer purchasing decisions, while Liu (2012) argued that numeric scores alone are insufficient without sentiment and context.  
  
Ferreira et al. (2016) demonstrated that machine learning models outperform traditional linear regression in retail demand forecasting, particularly in the presence of non-linear interactions. Taylor and Letham (2018) introduced Prophet, a scalable time-series model designed for business forecasting, though this method was not applicable in the current project due to the lack of time-based features.  
  
Real-world applications of similar analytics can be seen in Netflix’s recommendation system or Etsy’s pricing algorithms, both of which rely on a combination of behavioural, structured, and unstructured data.

## Methods – Data Collection

The dataset used was sourced from Kaggle and contained 2.2 million product entries, each with fields for:  
- Product identifiers (e.g. asin, title)  
- Price and rating (price, stars)  
- Review count (reviews)  
- Sales volume (boughtInLastMonth)  
- Category and bestseller status (categoryName, isBestSeller)

## Data Preparation and Cleaning

Initial inspection revealed no missing values, but significant skew and outliers were present in numerical features. Using the interquartile range (IQR) method, entries were filtered to exclude extremely high values in price, reviews, and boughtInLastMonth.  
  
\*\*Figure 1 Placeholder: Boxplot – Sales, Reviews, and Price Distributions\*\*  
\*This visual highlights extreme values and justifies removal thresholds to improve model reliability.\*

## Feature Engineering

To improve model performance and introduce domain logic, the following custom features were created:  
  
```python  
df['reviews\_per\_star'] = df['reviews'] / (df['stars'] + 1)  
df['price\_per\_star'] = df['price'] / (df['stars'] + 1)  
df['review\_intensity'] = df['reviews'] / (df['price'] + 1)  
```  
Each feature served a specific role:  
- reviews\_per\_star adjusts quantity of reviews by quality  
- price\_per\_star measures perceived value  
- review\_intensity captures engagement relative to cost  
  
\*\*Figure 2 Placeholder: Histogram – Engineered Feature Distributions\*\*  
\*Displays how new variables contribute to reducing skew and improving predictive signal.\*  
  
This step was particularly important because raw numerical features often fail to capture the nuanced relationships between product attributes and consumer behaviour. By introducing ratios, we allowed the models to generalise better across products of varying prices and review volumes. For instance, two products may have 1,000 reviews each, but if one has a much lower star rating, its perceived trustworthiness may be lower. Similarly, normalising price against rating provides a proxy for value-for-money, a critical factor in purchasing decisions. Feature engineering not only enriched the dataset, but also supported the theoretical basis for model predictions.

## Findings from Analysis

A correlation matrix was computed to evaluate relationships between variables. `isBestSeller`, `reviews`, and `review\_intensity` showed the strongest positive correlations with sales, albeit still weak overall.  
  
\*\*Figure 3 Placeholder: Heatmap – Correlation Matrix\*\*  
\*This visual supports the shift toward non-linear modelling due to weak linearity.\*  
  
\*\*Figure 4 Placeholder: Scatterplot – Reviews vs. Sales\*\*  
\*Reveals that high review counts do not guarantee high sales, supporting use of tree-based models.\*

## Modelling and Evaluation – Linear Regression

Used as a baseline, the linear model was trained on scaled features with the log-transformed target.  
  
```python  
from sklearn.linear\_model import LinearRegression  
model = LinearRegression()  
model.fit(X\_train, y\_train)  
```  
  
Performance:  
- MAE: 11.56  
- RMSE: 87.13  
- R²: -1.087  
  
\*\*Figure 5 Placeholder: Regression Plot – Predicted vs Actual (Linear Model)\*\*  
\*The model consistently underpredicted sales, confirming its limitations.\*

## Modelling and Evaluation – Random Forest Regression

The Random Forest model handled non-linearity and feature interactions more effectively.  
  
```python  
from sklearn.ensemble import RandomForestRegressor  
rf = RandomForestRegressor(n\_estimators=100, random\_state=42)  
rf.fit(X\_train, y\_train)  
```  
  
Performance:  
- MAE: 11.38  
- RMSE: 59.66  
- R²: 0.021  
  
\*\*Figure 6 Placeholder: Prediction Accuracy Plot – Random Forest\*\*  
\*Shows tighter clustering around actual sales values, though with considerable residual spread.\*  
  
While the MAE and RMSE were notably improved in comparison to the linear model, the R² value remained low. This reflects the inherent limitations of the dataset, which lacked external context such as marketing data, seasonal variation, or historical customer behaviour. However, Random Forest's ability to handle multicollinearity and capture non-linear interactions made it a better fit for this type of structured, categorical-heavy dataset. Hyperparameters such as the number of estimators were kept at default due to time constraints, but future iterations could benefit from cross-validated tuning to optimise performance further. Feature importance rankings could also provide insight into which attributes were most impactful.

## Discussion and Next Steps

The regression models applied in this project highlighted the challenge of predicting consumer behaviour based solely on structured product features. While Random Forest achieved improved error metrics over linear regression, the low R² scores indicate that the models failed to capture much of the variance in sales. This is likely due to unobserved external factors — such as seasonal demand, brand loyalty, and advertising — which were not present in the dataset. Nonetheless, variables such as review intensity and bestseller status emerged as useful business signals, suggesting that consumer engagement and product visibility hold predictive value.  
  
To enhance future iterations of this project, several directions are proposed. Firstly, natural language processing (NLP) could be applied to customer reviews to extract sentiment and thematic insights. Secondly, incorporating time-based variables would enable forecasting of seasonality and product lifecycle stages using models such as Prophet or ARIMA. Thirdly, ensemble methods such as XGBoost and LightGBM could be tested to potentially increase predictive accuracy. Finally, combining structured product metadata with behavioural data (e.g., customer clickstream or dwell time) would likely yield a more comprehensive understanding of online purchasing patterns.

## Conclusion

Although the models developed in this project demonstrated limited predictive power, the analysis uncovered valuable business insights. Bestseller status and review engagement were shown to be meaningful indicators of demand. The modelling exercise also reinforced the importance of incorporating additional data sources — such as sentiment and time — to improve accuracy and relevance. These insights have the potential to guide pricing strategies, promotional targeting, and inventory planning for e-commerce sellers.

## References

Brynjolfsson, E., Hu, Y.J. and Smith, M.D., 2003. \*Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers\*. MIT Sloan School Working Paper.  
Chen, P.Y., Wu, S.Y. and Yoon, J., 2004. \*The impact of online user reviews on product sales: The moderating role of product and consumer characteristics\*. Journal of Marketing Science.  
Liu, B., 2012. \*Sentiment analysis and opinion mining\*. Morgan & Claypool Publishers.  
Ferreira, K.J., Lee, B.H.A. and Simchi-Levi, D., 2016. \*Analytics for an online retailer: Demand forecasting and price optimisation\*. Operations Research.  
Taylor, S.J. and Letham, B., 2018. \*Forecasting at scale\*. The American Statistician.